PB87-229654

Environmental Methods Testing Site Project
Data Management Procedures Plan

Northrop Services, Inc./Environmental Sciences
Dayton, OH

Prepared for

Environmental Monitoring Systems Lab.
Research Triangle Park, NC

Jun 87

U.S. Department of Commerce
National Technical Information Service
NTIS

PB87-229654

EPA/600/4-87/023

June 1987

ENVIRONMENTAL METHODS TESTING SITE PROJECT Data Management Procedures Plan

by

E.H. Barrows
Northrop Services, Inc. - Environmental Sciences
2 Triangle Dirve
Research Triangle Park, NC 27709

Contract 68-02-4444

EMTS Project Coordinator
R.H. Jungers
EMTS Data Management Project Officer
T.C. Lawless
Monitoring and Assessment Division
Environmental Monitoring Systems Laboratory
Research Triangle Park, NC 27711

ENVIRONMENTAL MONITORING SYSTEMS LABORATORY
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
RESEARCH TRIANGLE PARK, NC 27711

REPRODUCED BY
U.S. DEPARTMENT OF COMMERCE
NATIONAL TECHNICAL
INFORMATION SERVICE
SPRINGFIELD, VA. 22161

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA/600/4-87/023	2.
3. RECIPIENT'S ACCESSION NO. PB87 229654/AS	
4. TITLE AND SUBTITLE: ENVIRONMENTAL METHODS TESTING SITE PROJECT: Data Management Procedures Plan	5. REPORT DATE: June, 1987
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S): Edward H. Barrows	8. PERFORMING ORGANIZATION REPORT NO.: Northrop Services, Inc.
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Northrop Services, Inc. 2 Triangle Drive Research Triangle Park, NC 27709	10. PROGRAM ELEMENT NO.: DU L-104, Project F-16,01
	11. CONTRACT/GRANT NO.: 68-02-4444
12. SPONSORING AGENCY NAME AND ADDRESS: Environmental Monitoring Systems Laboratory Office of Research and Development U.S. Environmental Protection Agency Research Triangle Park, NC 27711	13. TYPE OF REPORT AND PERIOD COVERED
	14. SPONSORING AGENCY CODE: EPA-600/08
15. SUPPLEMENTARY NOTES	

16. ABSTRACT

The Environmental Methods Testing Site (EMTS) Data Management Procedures Plan identifies the computer hard-are and software resources used in the EMTS project. It identifies the major software packages that are available for use by principal investigators for the analysis of data. It establishes standard methods for the transfer of data sets between the two principal operating locations. The procedures to add data to the EMTS data base, to extract data, and to generate reports are detailed. The responsibilities for the controlled, orderly and efficient use of the resources are delineated.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Environmental Methods Testing Site Data Management NCC Hardware Configuration Geographic Information System ARC/INFO Statistical Analysis Processors File Transfer Procedures		

18. DISTRIBUTION STATEMENT: RELEASE TO PUBLIC	19. SECURITY CLASS *(This Report)*: UNCLASSIFIED	21. NO. OF PAGES: 46
	20. SECURITY CLASS *(This page)*: UNCLASSIFIED	22. PRICE

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

DISCLAIMER

The information in this document has been funded wholly or in part by the United States Environmental Protection Agency under Contract 68-02-4444 to Northrop Services, Inc. – Environmental Sciences. It has been subject to the Agency's peer and administrative review, and it has been approved for publication as an EPA document.

Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

ABSTRACT

The Environmental Methods Testing Site (EMTS) Data Management Procedures Plan identifies the computer hardware and software resources used in the EMTS project. It identifies the major software packages that are available for use by principal investigators in the analysis of data. It establishes standard methods for the transfer of data sets between the two principal operating locations. The procedures to add data to the EMTS data base, to extract data, and to generate reports are detailed. The responsibilities for the controlled, orderly, and efficient use of the resources are delineated.

This report was prepared under Contract 68-02-4444 by Northrop Services, Inc. – Environmental Sciences under the sponsorship of the U. S. Environmental Protection Agency. This report was completed on June 15, 1987.

CONTENTS

Abstract ---------- iii
Figures ---------- vi
Table ---------- vi
Abbreviations ---------- vii
ix

1. Introduction ---------- 1
2. Data Processing Configurations ---------- 3
   - EMSL/RTP ---------- 3
     - Hardware ---------- 3
     - Mainframe computer at RTP ---------- 4
     - Software and data structures ---------- 5
   - EMSL/LV ---------- 10
     - Hardware ---------- 10
     - The VAX 11/785 and peripheral hardware ---------- 11
     - Software and data structures ---------- 13
   - Communications Hardware Configuration ---------- 15
3. Responsibilities and Procedures ---------- 18
   - Criteria and Procedures for Accepting Data ---------- 18
   - EMTS Data Set Inventory System ---------- 20
   - Tabular Data – EMSL/RTP ---------- 21
     - Site-characterization information ---------- 22
     - EMTS field study information ---------- 24
   - Cartographic Data – EMSL/LV ---------- 25
     - Site characterization information ---------- 26
     - EMTS field study information ---------- 27
   - Data flow (Communications) ---------- 28
     - Flow of data into the EMTS data base ---------- 29
     - Flow of data from the NCC to the GIS System ---------- 30
     - Flow of data from the EMTS data base to the user ---------- 30
   - Data Security ---------- 31
     - Access control ---------- 31
     - Backups ---------- 31
     - ERC library documentation ---------- 31

References ---------- 32

Appendix

A. Detailed System Diagrams ---------- 33

Preceding page blank

FIGURES

Number		Page
1	Configuration of the NCC IBM 3090-200	4
2	Configuration of the ISDAS system in Las Vegas	11
3	Sources of data in various formats entered into the ISDAS System in Las Vegas	12
4	EPA communication network diagram	16
5	Generalized view of the data flow through the EMTS System	19
A-1	Detailed diagram of the overall NCC communication network	34
A-2	Detailed diagram of the Region 4 communications network	35
A-3	Detailed diagram of the EMSL/LV communications network	36
A-4	Detailed schematic of the NCC IBM 3090/200 mainframe computer	37
A-5	Detailed diagram of the ISDAS	38

TABLE

Number		Page
1	TYMNET Access Numbers	17

ABBREVIATIONS

ADABAS	-- Adaptable DAta BAse System
BMDP	-- Biomedical Computer Programs
BPI	-- Bits Per Inch (referring to tape density)
DBMS	-- Data Base Management System
DMA	-- Direct Memory Access
EDSIS	-- EMTS Data Set Inventory System
ELAS	-- Earth Resources Laboratory Analysis Software
EMSL	-- Environmental Monitoring Systems Laboratory
EMTS	-- Environmental Methods Testing Site
ERC	-- Environmental Research Center
ESP	-- Economic Software Package
ESRI	-- Environmental Systems Research Institute
GIS	-- Geographic Information System
IDMS	-- Integrated Database Management System
IMSL	-- International Mathematical and Statistical Library
ISDAS	-- Integrated Spatial Data Analysis System
LPM	-- Lines Per Minute
LV	-- Las Vegas
MSS	-- Multispectral Scanner Data
NCC	-- National Computer Center
NDPD	-- National Data Processing Division
PC	-- Personal Computer
QA	-- Quality Assurance
RTP	-- Research Triangle Park, North Carolina
SAS	-- Statistical Analysis System
SMSA	-- Standard Metropolitan Statistical Area
SPSS	-- Statistical Package for the Social Sciences
TSP	-- Time Series Processor
UNLV	-- University of Nevada at Las Vegas
USGS	-- United States Geological Survey
VAN	-- Value-Added Network

SECTION 1

INTRODUCTION

The purpose of the Environmental Methods Testing Site (EMTS) project is to advance the state of the art of assessing human exposure to environmental pollutants. The specific objective of this project is to provide a well-characterized site in which to develop, test, and compare multimedia exposure methodology. Characterization of this site requires a well-managed computerized data base to store and efficiently retrieve data. Detailed characterization of one site, rather than less complete characterization of several sites in separate cities, will enable investigators to conduct a number of exposure assessment studies more quickly and economically. The approach to fulfulling the purpose and objective of the EMTS project is to set up a single, well-characterized site, and then invite principal investigators to conduct relevant field studies at this site.

The EMTS project has selected the Chattanooga Standard Metropolitan Statistical Area (SMSA) for this purpose. The SMSA contains the city of Chattanooga, TN, and the six counties surrounding it. There are three counties in Tennessee, and three in Georgia.

The overall output goals of the EMTS project are to provide computer graphics and analysis and terminal views of data. These goals are further defined below.

1. Providing a common data base for users of the EMTS characterizing the site where the studies are performed.

2. Providing products that will assist users in the conduct and analysis of their studies. These products will in most cases be tools to facilitate various types of data analysis.

3. Providing products that will assist EPA in advancing methodologies for assessing human exposures.

This data management procedures plan is designed to provide direction to and coordination of those aspects of the EMTS project dealing with the acceptance, storage, transfer, manipulation, and validation of data. It provides guidelines for EPA and contractor personnel and assigns responsibilities to ensure data integrity and completeness. The data management procedures plan should be used in conjunction with other EMTS documentation[1-3] to create a clearer understanding of the goals, concepts, objectives, procedures, and responsibilities involved in the project.

The EMTS data management concepts and procedures outlined in this document are intended for the use of all persons and agencies that may be in a position to control, process, store, or otherwise manipulate data in the EMTS. It is primarily intended to function as a guide for coordinating personnel at EPA's Environmental Monitoring Systems Laboratory at Research Triangle Park, NC (EMSL/RTP), and at Las Vegas, NV (EMSL/LV), and the contractor personnel supporting these two laboratories. It is not intended as a guide for EMTS principal investigators, or for use by other persons or institutions who contribute data to the EMTS data base.

Quality assurance (QA) of the EMTS data is an integral part of this data management guide. Therefore, this document refers to the concepts and philosophies of good QA procedures throughout, rather than in a separate section. The EMTS Quality Assurance Program Plan[2] should be referenced for further information on this subject.

The EMTS data base, for the purposes of this document, is defined as the collection of digitized data that have been through the QA process and are available for use by the EMTS project personnel.

The Plan is divided into two major sections. The topics addressed in these sections are described below.

Section 2 of this document, Data Processing Configuration, contains a discussion of the resources, both in terms of hardware and software, that are available for the handling, storage, manipulation, and reporting of data. Part of this section deals with the facilities and software available in the EPA National Computer Center (NCC) at RTP, the computer center through which most data will enter the EMTS system. Also in Section 2 is a discussion of the hardware and Geographic Information System (GIS) that are installed at the EMSL/LV site. Much of the data will flow from the NCC to this site for processing into spatial representations, or maps. Lastly, Section 2 contains a description of the data transmission (communications) facilities available for the EMTS.

Section 3 of this plan, Responsibilities and Procedures, deals with the assignment of responsibilities and procedures to be followed for data management in the EMTS. The EMTS Data Set Inventory System (EDSIS) is discussed. This is followed by a section describing procedures for handling tabular data and another describing procedures for spatial or cartographic data. These sections are, in turn, each divided into two identical subsections, "Site-Characterization Information" and "EMTS Field Study Information." Each of these subsections is divided into two further breakdowns, "Digital Data" and "Nondigital Data." Section 3 continues with a description of the anticipated data flows among EMSL/RTP, EMSL/LV, and the principal investigators. Finally, data security is addressed. An appendix contains detailed illustrations of the EMTS systems and networks.

SECTION 2

DATA PROCESSING CONFIGURATIONS

This section addresses the hardware and packaged software that are available to the users and managers of data in the EMTS System for the storage, processing, and transfer of data. These discussions are organized with an explanation of the hardware resources available for data collection and manipulation followed by a discussion of the software resources available for the analysis of data sets.

The principal resources available to the EMTS project are situated in two geographic locations. One facility is located at the Research Triangle Park in North Carolina, the other in Las Vegas, Nevada. The resources of both locations operate together to facilitate the goals of the EMTS.

The general flow of data in the EMTS will normally be directed from various originating sources into the NCC IBM 3090/200 computer at RTP for data collection and initial manipulations. These data will be stored in a relational data base management system (DBMS), and from these data, customized subsets will be extracted for analysis and manipulations. Analysis resources of two principal types are available for the EMTS project: (1) statistical packages on the IBM mainframe may be used for analyzing tabular data sets, and (2) the GIS available on the DEC VAX 11/785 at the EMSL/LV facility may be used for analyzing spatial data sets. Data transmission resources include tapes and on-line facilities to move files between computer systems.

EMSL/RTP

The Monitoring and Assessment Division of EMSL/RTP, with its associated support contractor, routinely processes large data sets from scientific research projects on the large IBM 3090-200 mainframe computer located there. This Division supports the EMTS project with the expertise and experience of its personnel working with the IBM system. The NCC resources enhance the ability of the EMTS principal investigators to make effective use of the site-characterization data for their projects.

Hardware

The National Computer Center is one of the largest, most modern, high-speed computer centers in the nation. Located within the U.S. EPA Environmental Research Center at RTP, its mission is to support EPA in both scientific and administrative applications. The NCC's physical facilities, hardware, and software constitute the central computer system, and a communications network allows distribution of computer services to remote locations.

The National Computer Center operates three principal data processing systems: an IBM 3090-200 (Figure 1), a Sperry 1100/82, and a minisystems operation consisting of three DEC VAX 11/780s, an IBM 4341, and two PDP 11/70s. The management of the NCC systems is organizationally structured within the National Data Processing Division (NDPD) of the Office of

Administration and Resources Management. NDPD personnel are responsible for managing the computer hardware, software, telecommunications network, and support services for this facility.

The hardware at the NCC facility accomodates a wide variety of input and output media. Punched cards, tapes, printed output, microfiche, and graphics are products of the services provided.

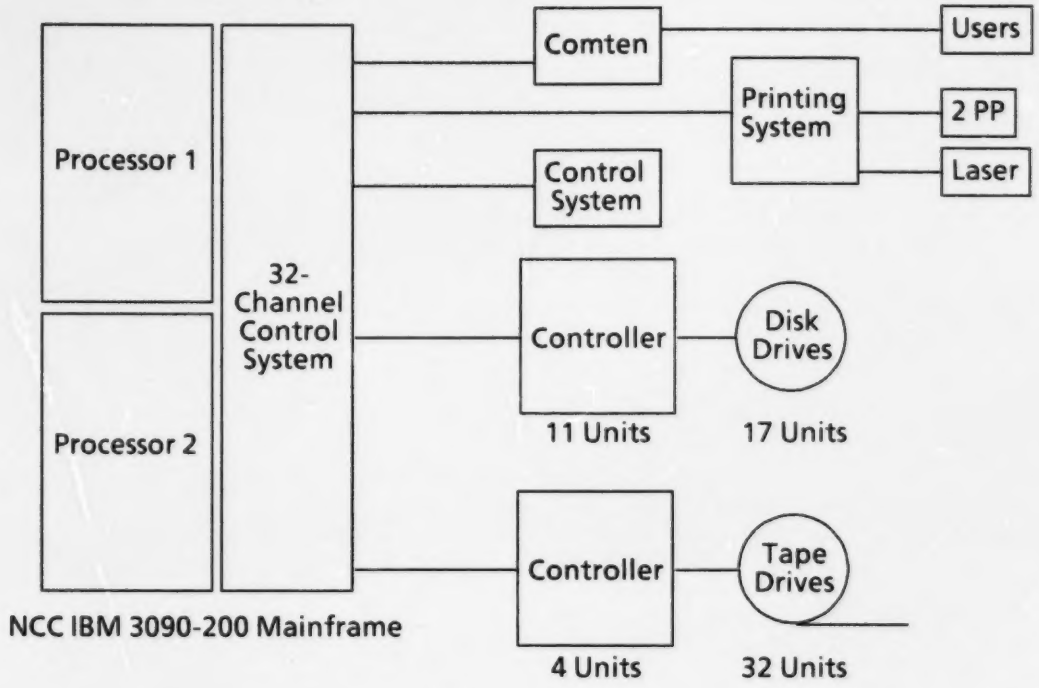

Figure 1. Configuration of the NCC IBM 3090-200.

Mainframe Computer at RTP

The collection of data processing devices that constitute a computer installation, along with the arrangement of those devices, is generally known as the hardware configuration. The NCC-IBM 3090/200 configuration consists of one processing unit which contains two integrated central processing units, the central storage, channel subsystems, channel paths, controllers, and input/output devices.

The processing unit is the controlling center. It provides hardware for the execution of programs, main storage addressing, and initiation of communication between main storage and the input/output devices. The processing unit of the NCC-IBM is an IBM 3090-200 and consists of two integrated central processors.

Central storage (memory) provides the processing unit with directly addressable, fast-access storage of data. Both data and programs must be loaded into main storage (from input devices) before they can be executed. The NCC-IBM configuration has 64 million bytes of directly addressable on-line memory, or central storage, on the IBM 3090-200.

Input and output operations are handled in the processor unit by the channel subsystem, which provides 48 channels, 32 standard and 16 optional. All channels can be configured for block-multiplexer operations, and as many as four channels can be configured for byte-multiplexer operation. Any channel not needed for byte-multiplexer operation can be configured for block-multiplexer operation.

The eight drum storage devices in the NCC-IBM configuration are used by the operating system for functions such as paging.

Two types of disk devices are used. The first type is the STC 8650 (analogous to the IBM 3350) dual density disk drives with a storage capacity of approximately 635 million bytes of information. These devices have 19,069 bytes per track, 30 tracks per cylinder, and 1110 logical cylinders per pack. The second type is the STC 8380 dual density disk drive, which has 1.26 billion bytes of information stored as 47,476 bytes per track, 15 tracks per cylinder, and 885 cylinders per pack. The 8380s not only offer greater capacity, but they also offer a 50% higher access speed and 2.5 times the transfer rate of the NCC's 8650 disks. The total on-line disk storage capacity is approximately 1.9 billion bytes of data.

The NCC-IBM 3090-200 utilizes three types of tape drive, nine-track single density, nine-track dual density, and seven-track multiple density. Nine-track tapes can be written and read at 800 bits per inch (BPI) single density, or at 1600 or 6250 BPI dual density. Seven-track tapes can be written and read at 200, 556, or 800 BPI (triple density).

Incremental plotters are available for graphic output. They consist of Model 925 memory controllers and Model 1051 digital incremental drum plotters manufactured by CalComp. The controllers include read-only tape drives for nine-track, 1600-BPI tapes.

Two IBM 3211 printers each run at 2000 lines per minute (LPM). This printing capacity is augmented by the high-volume, high-quality output of an IBM 3800 laser printer, which runs at 20,040 LPM. In addition, an IBM 3505 card reader (1200 cards per minute) and an IBM 3525 card punch (300 cards per minute) are included in the system.

Software and Data Structures

The NCC-IBM is equipped with a wide range of software packages to facilitate file and data manipulations. These packages encompass DBMS, econometrics, mathematical modeling, plotting and graphics, retrieval and reporting, and simulation and statistical analysis. Of particular interest to the EMTS are two types of software packages, (1) a DBMS that is used for storing and manipulating the EMTS data and (2) statistical analysis packages that can be used to analyze data sets selected from the DBMS.

Several DBMSs are available on the NCC-IBM mainframe computer. These include ADABAS (the Adaptable Data Base System), the Integrated Database Management System (IDMS), and System 2000. In addition, a DBMS is included in some software packages, such as Statistical Analysis System (SAS) and FOCUS.

Selecting the optimum DBMS for the EMTS from among these available systems requires a great deal of analysis of the potential uses of the system. Trade-offs must be made in the selection of a system, some of which are not entirely technical. The designated standard EPA DBMS is ADABAS; however, use of the system is restricted to specific applications, and it is used for routine processing of previously defined records in a rigid structure. The EMTS will use some data sets that are already part of the ADABAS environment, and ADABAS certainly has the capacity to handle the number of users and data anticipated for the EMTS project. However, the ADABAS administrative control

structure places severe restrictions on the types of data and users that can utilize this system. It is not suitable for use in the EMTS because it is not administered in a manner that will allow the degree of flexibility required with new types of data.

The IDMS requires a high degree of experience and training for users, thereby limiting its usefulness as a DBMS with the potential for many users. System 2000 utilizes an inverted hierarchical structure, which although relatively easy to use, lacks the flexibility of a modern DBMS. The data base management capabilities included in the SAS package utilize fixed-length records in a matrix organization. This arrangement works well for statistical analysis but is not as flexible as a relational model.

The data base manager selected for use with the Geographic Information System is INFO. This system is not available in a mainframe configuration and because of the large quantity of data expected to be stored in the EMTS would not be practical by itself. It will accept data from other data management systems.

Fortunately, the FOCUS DBMS is powerful enough for the quantity of data the EMTS data base will contain. It also is easily used for retrievals, is flexible, and supports a relational data structure. FOCUS does not have the same type of rigid controlling procedures accompanying it as ADABAS does. The only disadvantage of using FOCUS is the need to duplicate some data already in the ADABAS data base.

FOCUS is a comprehensive information processing system designed to run interactively on the IBM 3090-200 mainframe, as well as on IBM personal computers (PCs) and compatible units. It contains features for describing external files, as well as FOCUS files, for generating reports. It provides a dialogue manager, incorporates full-screen processing capabilities that can prompt for data entry, and performs data validation based on user definitions. In FOCUS, most conventional programming language structures are replaced with nonprocedural English sentences, in an attempt to avoid cryptic, computer-inspired notations. FOCUS is powerful enough to build menu-driven applications yet easy enough for nontechnical end users to perform their own ad hoc queries and reports.

The capabilities of FOCUS are best viewed in terms of the application processes, divided into the four following components.

1. On-line data entry, editing, and updating of master data bases as a single, continuous process.

2. On-line data entry, editing, and creation of transaction files, followed by batch updating of master data bases as a separate process.

3. Sub-data base creation, which consists of extracting records and/or fields from a production data base(s) and storing them in a data base structure for back-end processing. Sub-data base creation may involve transmission of files to or from PCs and from the NCC mainframe to other mainframes or minicomputers.

4. Back-end processing, which primarily consists of retrieval, presentation, and analysis of data maintained in the transaction files, master data bases, or sub-data bases. Back-end processing may involve canned (routine) processing and ad hoc activities. Canned processing is a standardized procedure that can be performed by executing a few commands. Typically, canned programs are repetitive or commonly used functions, for example, standard, routine reports and queries.

FOCUS is an integrated information control system that provides a nonprocedural, English-like command language and a DBMS. Its three major capabilities are retrieval and analysis, data base structure and storage, and data maintenance. Retrieval and analysis provides the following capabilities.

- FOCUS and PC/FOCUS allow access to FOCUS and non-FOCUS files. FOCUS can retrieve information from ADABAS, System 2000, IDMS, IMS, VSAM, ISAM, Sequential, and FOCUS files.

- A simple set of English-like commands, in concert with master file descriptions, allows quick generation of tabular reports and graphs.

- The menu-driven, dialog-oriented, interactive statistical programs provide a comprehensive statistical analysis capability.

- The graphics modules allow creation of simple business graphics, such as bar charts, pie charts, and histograms.

For its data base structure and storage capabilities, FOCUS is based on a shared, relational-structure data model. This model supports both simple and complex applications with small or large volumes of data. The data base facility also provides a dynamic file-joining (coupling) and logical file views. In addition, data encryption and security at the field-value level are available for FOCUS files. For its data maintenance capability, FOCUS has a full-transaction processing facility, including a data base editor, 3270 full-screen processing, and central data base control for multi-user access and processing.

The developers of FOCUS have extended the capabilities of the system to the IBM PC/XT and PC/AT microcomputers. PC/FOCUS, the personal computer version, is compatible with the mainframe version. This compatibility includes the capability to transfer application code and sequential files to the PC and then to return them for execution on the mainframe and vice versa.

FOCUS has introduced the concept of "talk" technology, which allows the most novice user to become quickly productive through the use of menu-driven questions that appear on a full-screen terminal. For example, the user may access TABLETALK to develop and execute reports and FILETALK to create data base file structures. The "talk" technology operates on both the mainframe and PC-based versions of FOCUS.

Regardless of the DBMS used for storage and manipulation, statistical analysis of the data is usually best performed by one or more of the statistical packages available on the NCC-IBM 3090-200 mainframe computer. The selected data sets can be the results of merging and collating data from a variety of sources and data bases. The following statistical analysis programs are available on the NCC-IBM computer.

Statistical Analysis System--

Statistical Analysis System (SAS) is an integrated system for data management and statistical analysis. It combines statistical routines, plotting, data manipulation, and report-writing capabilities.

SAS has a PL/1-like language that is used to transform or delete variables, create new variables, create data sets, merge data sets, and print all or part of a data set. SAS accepts data in most forms

and can be accessed through many input devices. The system also performs the following general data manipulation and many statistical functions.

- Linear and nonlinear regression
- Analysis of variance and covariance
- Matrix manipulation
- Multiple and partial correlation coefficients
- Spearman's and Kendall's correlation coefficients
- Frequency and cross-tabulation tables
- Assorted printed and plotted output

Statistical Package for the Social Sciences--

Statistical Package for the Social Sciences (SPSS-X) is an integrated system of computer programs for the analysis of social science data. The package includes the following capabilities in addition to descriptive statistics, simple frequency distributions, and cross-tabulations.

- Simple correlation
- Partial correlation
- Means and variances for stratified subpopulations
- One-way and n-way analysis of variance
- Multiple regression
- Discriminant analysis
- Scatter diagrams
- Factor analysis
- Canonical correlations
- Guttman scaling

The SPSS-X package has facilities to generate new variables that are mathematical and/or logical combinations of existing variables; to record variables; and to sample, select, or weight specified cases.

Economic Software Package--

The Economic Software Package (ESP), a revised and extended version of an earlier system called Time Series Processor (TSP), is a computer language for the statistical analysis of time series by the ordinary least-squares method or the two-stage least-squares method. It is designed to carry out the computational and data processing steps that occur routinely in economic research. Input to ESP

is made through the system's own free-format language. The following features are included in the ESP package.

- Fitting and extrapolation
- Matrix manipulation
- Time series plots
- Statistical measures
- Normalized random number generator
- Scatter diagrams
- Exponential trend lines

International Mathematical and Statistical Library--

The International Mathematical and Statistical Library (IMSL) is an extensive collection of mathematical and statistical subroutines and groupings, which are listed below.

- Categorized data analysis
- Eigenvalue analysis
- Forecasting
- Econometrics
- Analysis of experimental design data
- Basic statistics
- Time series analysis
- Generation and testing of random numbers
- Interpolation
- Approximation
- Curve fitting
- Linear algebraic equations
- Sampling
- Utility functions
- Vector and matrix arithmetic
- Linear programming

Biomedical Computer Programs--

The Biomedical Computer Programs (BMDP) are a comprehensive and integrated library of statistical programs. These programs may be loosely classified into six series: data description, frequency tables, regression analysis, analysis of variance, multivariate analysis, and special analysis (which includes nonparametric statistics and cluster analysis).

BMDP offers the user a broad set of descriptive and analytical procedures. These procedures have convenient defaults, offer supplementary printer graphics, and have comprehensive documentation. Data grouping or selection may be changed through the control language. The generation of new variables, as well as data modification, can be accomplished through the use of a control language or the inclusion of a FORTRAN subroutine in the BMDP run. The scope of the program includes the following features.

- Univariate statistics (means, skewness, and histograms)
- Bivariate f and t tests and comparison of cells

- Scatter plots and probability plots
- Missing values estimates, patterns, and correlations
- Multiple linear, stepwise, best subset, principal component, and polynomial regression, as well as nonlinear regression
- Analysis of variance of fixed-effect or mixed models, which include repeated measure and unequal cell size options
- Nonparametric statistical analysis
- Factor analysis, canonical correlation, discriminant analysis
- Cluster analysis on case, variables, or blocks

Executions of the statistical analysis packages outlined above is one form of output from the NCC. These analyses may be performed by the principal investigator or other parties, who should first utilize dummy data to verify the integrity of the data management procedures. The second form of output from the NCC will be files of selected subsets of the data base that are prepared and transmitted to the Las Vegas-based GIS for conversion into spatial representations of the data.

EMSL/LV

The Environmental Monitoring Systems Laboratory at Las Vegas provides the EMTS project with the capability to perform a variety of spatial analyses of environmental data. Experts in remote sensing and GIS work in a highly sophisticated hardware and software environment to provide an entirely new way of investigating environmental problems. GIS analyses provide mechanisms for linking relational data base functions with a computer mapping system, allowing locational variables and relationships to become evident. More importantly, GIS provides the capability for models to interact with the data and for data layers to interact with each other in a common framework.

The EMSL/LV will receive subsets of the relational data bases available from the NCC, digitize map data, and procure spatial data already in digital format. These data sets will be entered into the Integrated Spatial Data Analysis System (ISDAS) and analyzed by using the GIS to provide maps, files, and statistics that will enable EMTS investigators to assess environmental monitoring methods.

Hardware

Two separate computer systems are available at EMSL/LV. The primary Laboratory computer is a VAX 11/785. This machine is linked to the NCC mainframe through several 9600-baud, dedicated lines. A variety of communications ports connect users throughout EMSL/LV to this system, and through the system to the NCC IBM. EMTS data sets may be passed directly to EMSL/LV through the RTP-to-LV link. The Laboratory VAX computer provides several data bases and a variety of statistical analysis packages for general scientific support.

The ISDAS is supported by two computers linked together with ETHERNET. The principal GIS unit is also a VAX 11/785, and the principal image-processing device is a Concurrent Computer Corporation (formerly Perkin-Elmer) Model 3210. Figure 2 outlines the basic ISDAS system configuration at the ERC facility in Las Vegas, and a more complete design schematic is included as Figure A-3 in Appendix A. The primary design criterion for the system was to provide maximum processing power for relatively few users, enabling them to handle large spatial data sets. The following discussion refers to the ISDAS and not to the Laboratory System, unless stated otherwise.

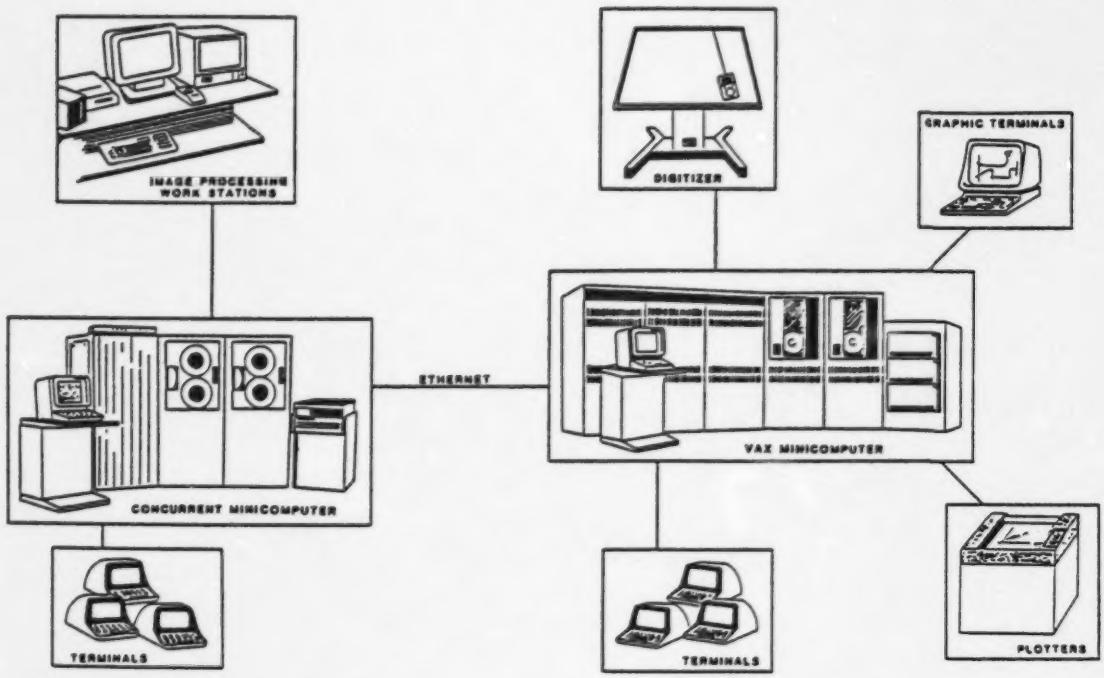

Figure 2. Configuration of the ISDAS system in Las Vegas.

The VAX 11/785 and Peripheral Hardware--

The ISDAS VAX is configured with 8 Mb of main memory, two 1600/6250-BPI tape drives, three RU-81 456-Mb disk drives, user control storage, and 16 asynchronous input lines. External communication is provided by one modem set at 300/1200 baud and one set at 2400 baud. All other communication ports are occupied by internal RS-232 connections at 9600 baud. A Floating Point Systems Model 5305 array processor is linked to the system for processing large matrices of data.

Data input hardware is illustrated by Figure 3. Much of the EMTS data is input from tape and is already in digital format (such as data from the NCC IBM mainframe, photogrammetric workstations, or multispectral scanner data). Data that are in analog (map) format are digitized using a CalComp 9100 digitizing tablet (48-in. x 60-in. surface) with a resolution of 0.003 in. For some special applications, image data from sources such as maps or aerial photographs can be captured by using dual video camera systems on the Aeronca VGS 300 + video digitizing and analysis system, processed on its IBM PC/AT system, and transferred to the VAX using the KERMIT file transfer protocol or floppy disks.

A Tektronix Model 4125 and a Model 4111 graphics terminals are used as workstations for digitizing, analyzing, and displaying vector-oriented graphics. Five DECMate III workstations are used as general purpose terminals for the VAX and can also serve as stand-alone word processors. An LA-100 terminal provides convenient hard copy. Several IBM PC/AT computers, including an ERDAS workstation and the VGS 300 + system, can also be used as VAX or Concurrent terminals, either with direct connections or through modem links. This allows easy file transfers via floppy disks or hard disks, to or from the VAX.

Raster data analysis is provided using an ERDAS IBM PC/AT System 2 connected to the VAX with an IEEE 488 interface. This system functions as both a stand-alone system and a workstation linked to the VAX. A 19-in. monitor provides raster display in a 512- x 512-pixel, three-image-plane, true-color format.

The results of data analysis may be output to tape or IBM-compatible floppy disk. If hard copy is desired, maps may be produced in color on either of two devices. The primary color plotter is a Versatec Model 7224 electrostatic color plotter. This instrument plots at a resolution of 200 dots per inch on either paper or plastic media of 24-in. width and any length desired. The secondary plotting device is a Tektronix Model 4692 color graphics copier, which outputs on 8.5-in. x 11-in. media and is a rotating-drum, ink jet device. The 4692 copier, connected to both the Tektronix 4125 and 4111 graphics terminals, provides color reproductions of the displays on paper or clear acetate.

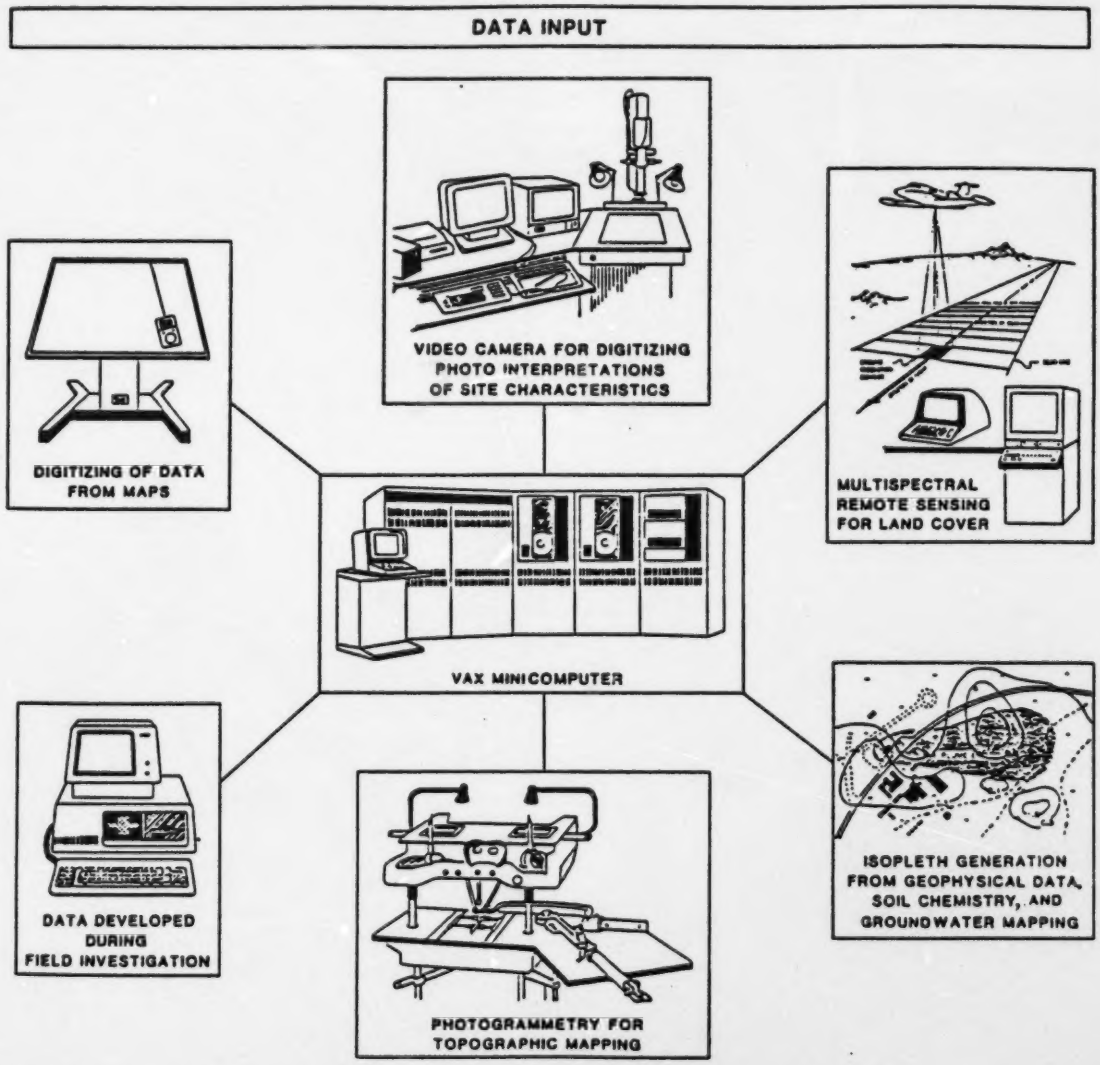
DATA INPUT
VIDEO CAMERA FOR DIGITIZING PHOTO INTERPRETATIONS OF SITE CHARACTERISTICS
DIGITIZING OF DATA FROM MAPS
MULTISPECTRAL REMOTE SENSING FOR LAND COVER
VAX MINICOMPUTER
DATA DEVELOPED DURING FIELD INVESTIGATION
ISOPLETH GENERATION FROM GEOPHYSICAL DATA, SOIL CHEMISTRY, AND GROUNDWATER MAPPING
PHOTOGRAMMETRY FOR TOPOGRAPHIC MAPPING

Figure 3. Sources of data in various formats entered into the ISDAS system in Las Vegas.

Image-Processing Hardware--

The principal image-processing machine is a Concurrent Computer Corporation Model 3210 minicomputer with 4 Mb of main memory in 32-bit words, Winchester 300 Mb fixed disk, 300 Mb of mountable disk (with multiple disk packs), two 800/1600-BPI tape drives, and a 300-LPM line printer. This central processing unit is linked to the VAX via an ETHERNET interface and shares the same room.

Raster (image) data, such as LANDSAT or SPOT satellite data or digital elevation models, are usually entered into the system through tape. Other examples of raster data entry are files transferred from an Aeronca VGS 300 + video graphics system and direct decommutation of high density digital tapes of airborne multispectral scanner data (MSS).

Analysis is provided through a Comtal 8000 display device and a DeAnza ID 1124 image-processing device. The Comtal is a 512- x 512-pixel, single-image-plane, pseudocolor display device; and the DeAnza is a 1024- x 1024-pixel, three-image-plane, true-color device. Both systems are connected to the Concurrent system as direct memory access devices through a selector channel and interface board. Additional image processing is available on the IBM PC/AT-based ERDAS and Aeronca VGS 300 + systems connected to the VAX.

Raster hard-copy output is provided by a Dunn Instruments camera system linked to the DeAnza display device and by the Versatec electrostatic color plotter.

Software and Data Structures

Although much of the software used by the ISDAS is available from commercial sources, the state of the art and technology are such that the system must be considered developmental. Certain portions of the software are more mature and could be considered operational; however, many of the functions of the software need considerable programmer support. Therefore, the interdisciplinary team that handles the EMTS analyses includes software engineers, mathematicians, and computer scientists, as well as disciplinary scientists. The overall environment for ISDAS is one of growth and development, in which new software and techniques for analysis are being developed constantly.

GIS Software--

ARC/INFO, a product available from Environmental Systems Research Institute, Inc. (ESRI), of Redlands, CA, is the primary GIS input and analysis tool. This software provides all the necessary functions for digitizing, editing, combining, and plotting GIS data. The software combines a relational data base language (INFO) available from Henco, Inc., with a computer graphics and GIS analysis software (ARC) to provide a complete package of GIS functions on the ISDAS VAX 11/785.

In ARC/INFO, maps are usually represented as a collection of points which describe points, lines, and polygons. This type of representation is called "vector." The software maintains the x,y coordinate pairs that describe these points in several files with specialized formats, and it maintains the attributes that relate to these points in INFO files. The program ARC relates the coordinate pair data to the attribute (thematic) data through internal ID numbers common to both sets of data files. INFO allows up to 10 files to be "related" in a relational chain, and each file may contain multiple attributes.

ARC/INFO files are topologically structured. This means that the spatial relationships between map features are defined. Points are the most simple case and are simply represented by coordinates and associated attributes (topographic elevation, pollutant levels, etc.). Lines are represented as arcs

that have "nodes" (starting and ending points) with a linked collection of points between them. Arcs may have attributes assigned to them (such as County Road 49), and they may have attributes that bound them on the left and right (the line forms a boundary between soil 1 and soil 2). Arcs that share nodes are identified and may form polygons. Polygons may contain up to 5000 arcs of 500 points each. Tables of attribute data are maintained in INFO for points, lines, and polygons.

The maps are organized into groups called "coverages." Point, line, and polygon coverages are usually kept on separate files, even though they represent the same theme (or subject). Additionally, the coverages may be divided into "modules" (or "tiles"), which are edge-matched portions of the data base. (For the EMTS data base, the matrix of 54 United States Geological Survey (USGS) 7.5-min quadrangles forms the tile structure of the digital data base.) The modules need not be rectangular, although that is usually the convention. Thus, analysis may focus on selected coverages (such as soils, roads, and streams) for selected modules (such as the Fairmont Quadrangle).

Interactive display and analysis are accomplished by using a combination of the programs ARC, INFO, TIN, and NETWORK. These programs form the basic tools for spatial analysis of vector data. ARC/INFO is a combination of more than 300 routines for handling digital cartographic data and spatially associated tabular data. It is designed as a generic toolbox for data entry, manipulation, analysis, and display. NETWORK is a series of procedures that manage, analyze, and display geographic networks, such as city streets, railroads, rivers, and utility lines. Functions include managing DIME files, interactive districting, optimum routing, and facility allocation testing. TIN is a set of routines for handling continuous data, such as digital terrain models of results from trend surface analysis. Functions include contouring, point perspective (3-D) display, watershed modeling, slope/aspect analysis, Thiessen polygon calculation, cut-and-fill analysis, and digital terrain model analysis. Additional trend surface analysis may be accomplished using the SURFACE II contouring software.

Plotting on the Versatec color plotter is accomplished through the use of ARCPLOT routines interacting with special interface software from USGS and the custom software written by EMSL/LV programmers.

Image-Processing Software--

Raster files are formatted differently from vector files. Each location on the image is represented by a cell or picture element (pixel) in which an eight-bit number resides. (The image is really a matrix of integers encoded to represent topographic elevation, spectral radiance, land coverage class, etc.) These files may be extremely large, depending on the resolution of each pixel. EMSL/LV has translator programs to convert data files of both Earth Resources Laboratory Analysis Software (ELAS) and ERDAS to the ARC-GRID format and to the ARC/INFO vector format.

Raster data are handled by a combination of software routines residing on both the Concurrent computer and the VAX. The National Aeronautics and Space Administration ELAS forms the mainstay of the EMSL/LV raster processing capability. The ELAS package has more than 200 linked analysis routines. EMSL/LV has added at least another 50 routines to the package for the processing of Aircraft MSS data and has extensively modified much of the basic software. The Concurrent system is the primary computer for the ELAS software, but a version of the ELAS containing a limited number of functions is implemented on the VAX.

Additional raster processing is available on the VAX through the ERDAS image-processing system and the ESRI routines GRID and GRID/TOPO. The ERDAS software resides both on the VAX and an IBM PC/AT that has been specially configured as a VAX workstation. The ERDAS software performs a variety of raster GIS and image-processing functions (including image display, enhancement, and classification), which augment the ELAS functions of the Concurrent computer.

ERDAS data files may be easily transferred to the ARC/INFO vector format, although it is useful in some cases to maintain the raster data structure.

GRID/GRIDTOPO is a software package developed by ESRI for raster GIS processing. It primarily forms the translator between other gridded data formats and the ARC/INFO, but also contains some special-purpose, raster data-handling routines for processing digital topographic data.

Programming Environment--

Software is usually generated in FORTRAN, which is supported on both the VAX and the Concurrent systems. Special-purpose routines are sometimes generated in C or PASCAL languages, which are supported on the VAX. Data from each source usually require some reformatting before they can be converted to an ARC/INFO file structure, and some data structures require extensive conversion and editing.

COMMUNICATIONS HARDWARE CONFIGURATION

There are several methods of communicating data between the various elements of the EMTS system. The optimum system chosen for any given transmission will depend upon a variety of factors. Where speed is a primary consideration, the optimum method of communication most likely will be via dial-up telephone lines. This method is, in most cases, the most expensive and is also subject to undetected errors in transmission. These errors, due to noise in the lines or other equipment problems, have been known to drastically affect the quality of the data, sometimes with unforeseen results.

EPA has installed a network of landlines that connect the IBM mainframe of the NCC in RTP with the EMSL/LV VAX 11/785. This network provides a high-speed (9600-baud) cross-country linkup. A hard-wired link between the main EMSL/LV VAX and the VAX that runs the GIS has been installed. NCC is networked to the Washington Information Center and other centers with high-speed links. Users can tie into these high speed links with normal 300-1200-2400-baud, dial-up links or through other local networks. A schematic diagram of the communications network is shown in Figure 4.

The most effective method of transferring data to the ISDAS is via 1600- or 6250-BPI tapes in either binary or ASCII format. Small data sets may be transmitted via modem at rates up to 2400 baud. However, this speed is not suitable for the transmission of maps output from the ISDAS because of the high volume of data required to complete a map at high resolution.

The IBM mainframe computer can be accessed through a network of 3270-type terminals. This service is provided to all the Regional Offices, EPA Headquarters, and various contractor sites and laboratories. This service is available to users on leased telephone circuits and dial via protocol converters that allow for ASCII TTY terminals to emulate full-screen 3270 type terminals. Detailed diagrams of the relevant portions of the network (see Figure 4) are shown in Figures A-1 and A-2 in Appendix A.

Telecommunication services are also provided to the Washington Information Center's IBM computer via mainframe-to-mainframe links. This allows for exchange of information between mainframe computers and provides a link to the user through either IBM mainframe.

A computer link is currently being installed in the Tennessee Department of Health and Environment Systems and Procedure Division, within the SMSA selected for the EMTS project, to provide a direct connection into the NCC telecommunications network. Local agencies and personnel should access the EMTS resources through this device.

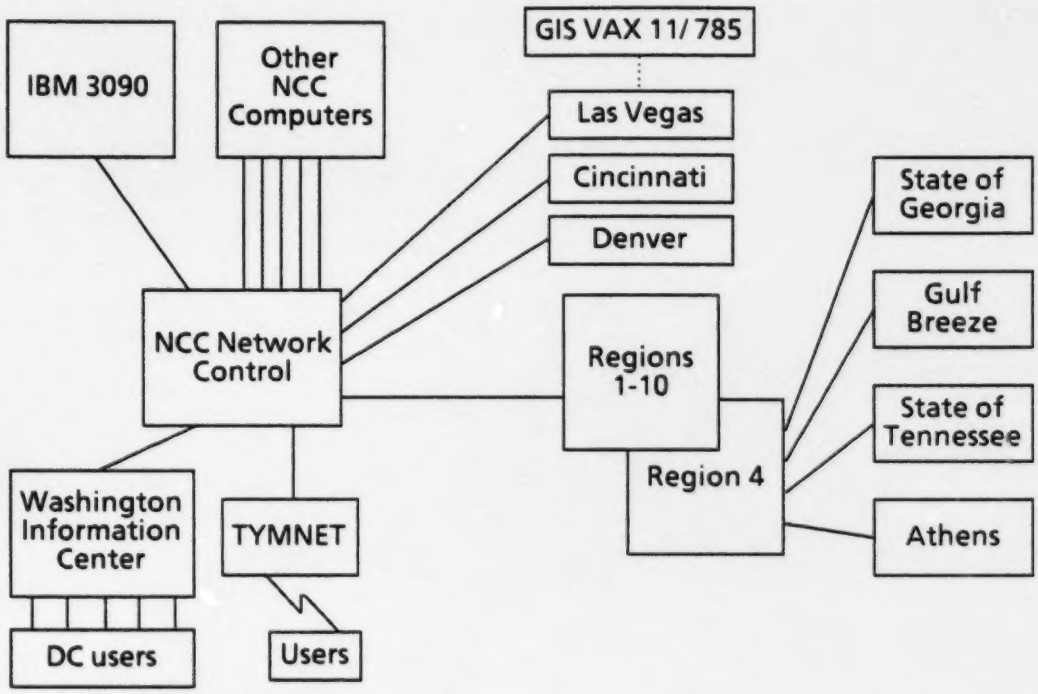

Figure 4. EPA communication network diagram.

Networking facilities for interactive access to the EPA IBM host computer and to other nodes of the network are also provided by TYMNET, a worldwide, Value-Added Network (VAN) supplier who, through a sophisticated network of dedicated circuits, nodal processors, local dial facilities, WATS lines, and automated network supervision, provides interactive access to hundreds of host computer systems. TYMNET "adds value" to conventional networking facilities in the form of improved efficiency, better economy, overall network supervision, enhanced routing, error detection and retransmission, local dial access in numerous cities and more.

TYMNET provides NCC with international and national access for interactive users. Users in areas where local numbers are not available are provided WATS numbers for access. Users in Washington, DC, RTP, regional offices, Las Vegas, and Cincinnati are provided with non-TYMNET facilities for access. Washington area users gain access through statistical multiplexers, and RTP users are provided direct access to the system.

TYMNET is constantly adding new facilities and increasing the list of access numbers, as shown in Table 1. Telephone numbers for access into the TYMNET network from other locations are maintained by the NCC Users Support, and users should call the NCC User Support at FTS 629-7862 or (800) 334-2405 to obtain the latest information on these TYMNET access numbers. Detailed procedures for user logons with a full-screen terminal are given in Appendix E of the *NCC IBM Users Guide*[4] available from NCC User Support.

TABLE 1. TYMNET TELEPHONE ACCESS NUMBERS

City	Number	Baud Rate
Chattanooga	615/265-1020	1200
Knoxville	615/690-1543	1200
Memphis	901/527-8006	1200
Nashville	615/855-3530	1200
	615/889-5790	2400
Atlanta	404-446-0270	1200
Norcross	404-446-1508	2400
Macon	912/746-2739	1200
Savannah	912/232-6751	1200

SECTION 3

RESPONSIBILITIES AND PROCEDURES

In this section of the EMTS Data Management Procedures Plan, the responsibilities and procedures for carrying out data management tasks in the EMTS are described. These responsibilities and procedures are applicable to the EMSL/RTP and EMSL/LV laboratories and their support contractors. These guidelines are designed to provide the structure to ensure an efficient, timely, responsive, coordinated, and accurate data management effort. It is anticipated that these procedures will be continuously augmented and updated by spoken and written communication between the principal action agencies. The procedures set forth here will evolve to become more detailed and comprehensive throughout the life of the EMTS program.

This section sets forth the procedures for accepting data for the EMTS data base. It contains a description of the EMTS Data Set Inventory System (EDSIS), which is a PC-based inventory of data sets available within the EMTS system. EDSIS is written, maintained, and operated by personnel at the Environmental Research Center (ERC) of the University of Nevada at Las Vegas (UNLV).

Procedures for handling the various forms of EMTS data are discussed next. The data that are to be stored, processed, and utilized for the EMTS project can be subdivided in three major ways: (1) tabular and cartographic data, (2) site-characterization information and EMTS field study information, or (3) digital and nondigital representations. Digital data, for the purposes of this plan, refer to data that are presently in a computer or computer-readable form, whereas nondigital data must be converted by keypunching, digitizing, or other data entry process to be read into a computer.

The transfer of data between the two principal locations (RTP and Las Vegas) of the EMTS project is a critical subject dealt with next in this section. Section 3 is concluded by a discussion of data security.

CRITERIA AND PROCEDURES FOR ACCEPTING DATA

Data may be accepted from any source that is capable of providing data relevant to the EMTS. However, before data are added to the data base, certain procedures must be followed and approvals obtained.

1. The data must be needed by and relevant to the EMTS project. The addition of new data sets that are not in the EMTS data base must be requested by the principal investigator through the Technical Review Subcommittee. Certain baseline and characterization data that are common to many studies have already been identified for inclusion in the data base by the UNLV ERC. The EMTS Steering Committee has the responsibility for approving studies, and thereby has indirect input to the determination of whether data are relevant and necessary.

2. In many cases, data may potentially be of value to more than one user. If the Project Coordinator, with technical guidance of ERC on a case-by-case basis,

determines the data set to be of global value to the EMTS project, and funding within the existing structure is sufficient, the data set(s) may be purchased, converted, processed, etc., with no additional financial support from the principal investigator. Data sets that probably would not be of global interest to the EMTS or for which a large expenditure of funds is required will not normally be purchased by the EMTS. Funding to obtain data sets that probably will be of interest to only one principal investigator will be decided on a case-by-case basis. In all cases, funding for the purchase of the data set(s), conversion of the data set(s), and processing for data refinement and/or analysis is ultimately the responsibility of the principal investigator requesting that a data set be added to the EMTS data base.

3. The EMTS data base contains only validated data. Quality assurance of the data must be verified by the QA specialist at the ERC in Las Vegas. No data may be entered into the EMTS data base that do not have a QA component or evaluation associated with it. The procedures for this step are outlined in the EMTS Quality Assurance Program Plan,[2] which should be referenced for further information.

A generalized view of the flow of data through the EMTS system is depicted in Figure 5. In this diagram, data are entered into the EMTS master data base from the QA function. This ensures that all data in the master data base have a quality assurance component. Sources of data may be other data bases that are identified in the site-characterization process.

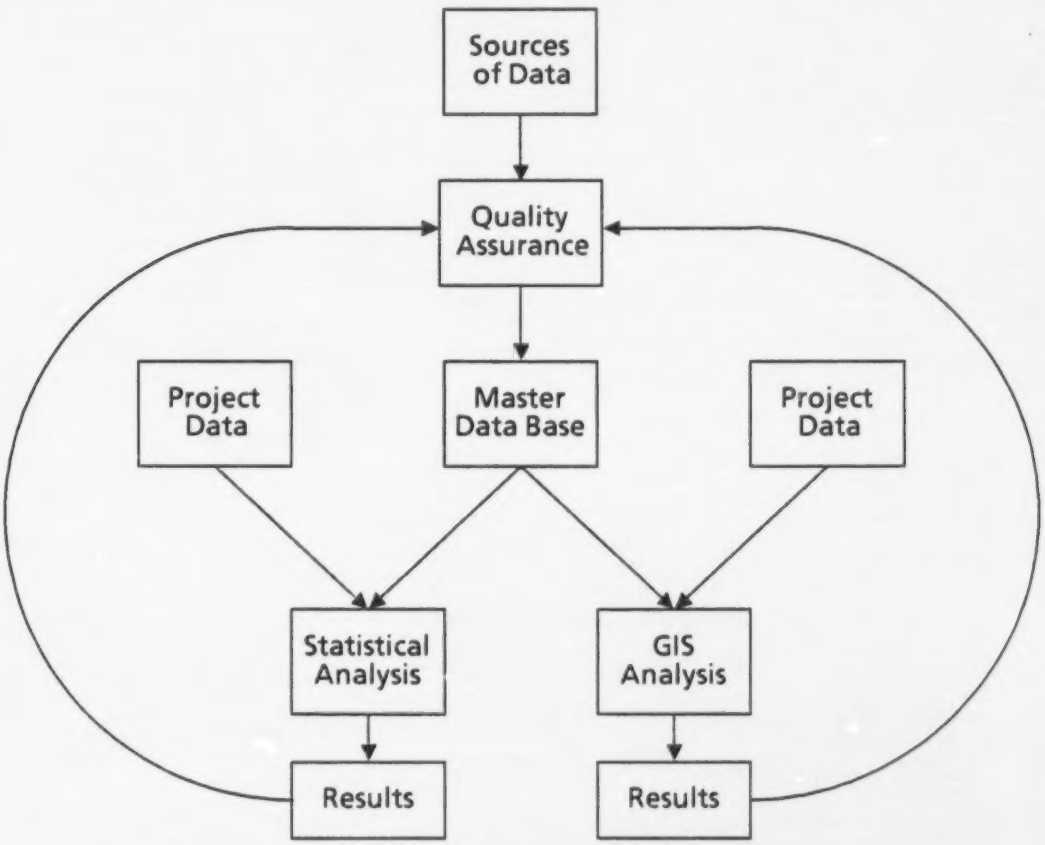

Figure 5. Generalized view of the data flow through the EMTS system.

Data will flow from the master data base into either or both of the analysis systems under the direction and control of a principal investigator. By either statistical or GIS analysis procedures, results of the work will be returned to the master data base for use by other researchers after approval by the QA specialist.

EMTS DATA SET INVENTORY SYSTEM

EDSIS, written and maintained by the ERC personnel at UNLV, is a compiled dBASE III questionnaire and reporting system designed to run on IBM and compatible PCs. Copies of the questionnaire-format program are sent to each participating agency or contractor that may be identifying data sets for possible inclusion in the EMTS data base. The program contains a provision for adding offices and agencies to the list of users of the inventory. These offices and agencies, in turn, fill out the automated questionnaire for each of the data sets that they consider relevant to the EMTS project and return it on a diskette to the ERC operator for addition to the master inventory. An instruction document supplied with the program identifies codes, formats, etc., to be used.

Before data are added to the EMTS data base, a description of the data set must be entered into the EDSIS and approved for inclusion in the system. The most recent copy of the completed inventory must be checked to avoid duplication. The approval step is based on the applicability of the data set to the overall EMTS goals and on QA documentation of the data. Each participating operating agency in the EMTS will receive a copy of the completed inventory each time the master copy at the ERC is updated. These procedures will ensure that duplicate data sets are not collected, and that adequate procedures have been followed for the data to be usable. The procedures will also ensure that all data sets in the EMTS data base are cataloged and are available to other users of the system.

The EMTS data library, maintained at the ERC, is a collection of hard-copy documents, maps, digital tapes, disks, and photographs. Most of these materials contain support information for the data sets within the EMTS data base. Support material is usually ancillary data or QC/QA reports. In some cases the entire data set is in hard copy only and is maintained in this library.

All materials stored within the library are referenced in the EDSIS. A status category for each data set listed in the file informs the user if the data set has been validated and approved for use in the EMTS data base, and if approved, (1) into which computer system it has been incorporated or (2) that the data set is available in the library only in hard-copy form.

Storage facilities for the library materials consist of locking file cabinets, a storage closet, map cases, dust-proof disk boxes, and metal tape storage cans. The films are stored in the closet within dust-proof canisters. Any special care needed for digital materials is provided when necessary.

All materials in the library are labeled with the agency identification number and reference number assigned during the inventory process. The library is arranged alphabetically by participating agency and then numerically by reference number.

Presently, the EMTS library contains a collection of various inventoried materials, including some nondigital data. All EMTS data that are validated and approved for use in the automated EMTS data base will be digitized and incorporated into the proper computer system. Data considered to be not relevant to the EMTS Project will be returned to the originating custodial agency for storage.

TABULAR DATA – EMSL/RTP

EMSL/RTP has the primary responsibility for identifying, locating, inventorying, collecting and processing data sets that are not exclusively spatial. (Spatial data sets can be represented only by a map and are handled by the EMSL/LV facility.) Tabular data sets have a series of values associated with one or more compounds, elements, etc., taken over a period of time at one or more locations. Examples of this type of data are values for pollutants that have been measured at sampling sites. These data are usually analyzed by statistical techniques. Summaries of this type of data are frequently presented in tables. On occasion, graphic portrayals of tabular data may be helpful to present an understanding of the data.

It is not the function of the EMSL/RTP office to determine which parts of the EMTS data may or may not be applicable to later studies. All tabular data sets that are accepted into and procured under the EMTS will be stored at the NCC facilities for possible use by other users. Under normal circumstances, they will be available to any future user of the system. Costs necessary for data extraction, processing, and printing or copying must be borne by the requesting principal investigator. The costs associated with access to the data on the NCC computer are also the responsibility of the principal investigators.

The principal investigator may request data in a format other than the format in which the data were first collected or added to the EMTS data base. Data sets will be customized by EMSL/RTP to meet the needs of the requester. For this purpose, the principal investigator must communicate directly with the EMSL/RTP data management personnel to define the data sets that are relevant to the particular study. The principal investigator may request that only selected records or selected parts of records be extracted from the EMTS data base. Multiple data bases may be accessed and extracted to create the required data set. EMSL/RTP personnel will then set up and execute the extraction program to create the data set(s) desired.

These customized data sets will be delivered to the requesting principal investigator. They will not be separately inventoried because this procedure would duplicate data already included in the EDSIS system. However, an index of the files that are thus created and distributed will be maintained at EMSL/RTP for historical documentation and to track data that have been sent to each principal investigator.

The data made available to principal investigators will be provided through the Project Coordinator by EMSL/RTP in the media and format requested. The preferred media for the selected data should be discussed by the principal investigator and the EMSL/RTP data management staff. EMSL/RTP personnel are experienced with, and the NCC provides the necessary hardware for, a range of options for the user. First, the principal investigator may obtain data from the NCC in sequential data files, which may be prepared on magnetic tapes with different coding structures (ASCII or EBCDIC), different densities (800, 1600, or 6250 BPI), and either seven or nine tracks for use on different computers. Special formats can be created if necessary. A second mechanism for the principal investigator to obtain data is through other transfer media that are available through PCs. IBM-formatted 5¼-in. diskettes at either 360 Kb or 1.2 Mb are currently in use and available. Other types and sizes may be created if necessary and if the required equipment is available.

A third mechanism for transferring selected data sets from the NCC to the principal investigator is use of an unprotected copy of the required data set created by EMSL/RTP personnel on the NCC IBM mainframe, which could be accessed on-line by a principal investigator through the NCC telecommunications network. Certain advantages may be gained by the principal investigator in using the data set through direct access techniques on the NCC IBM 3090-200 mainframe. The user is assured of the most recent version of the data set available. The statistical analysis resources of the packages available on the IBM, which are extremely powerful and comprehensive, may be a

significant help in understanding data. Data sets may be accessed by more than one individual; thereby the same resource is shared, resulting in a lower cost per user. The NCC telecommunications network is nationwide; therefore, data that are used and prepared at one location are also available at another site without a delay in transfer or the risk of having data lost or garbled in transmission.

Procedures detailing the steps a principal investigator should take to become a user of the NCC computer are beyond the scope of this plan. However, guidance in the process, appropriate forms, and approval can be obtained through the EMSL/RTP ADP Coordinator at FTS 629-2291 or (919) 541-2291.

Site-Characterization Information

During the site-characterization phase of the EMTS, tabular data will be evaluated, collected, stored, and processed on the NCC IBM 3090-200 mainframe computer by EMSL/RTP personnel. The data collected in this phase of the project will consist of data sets that have been assembled for a primary purpose other than the EMTS project. It is anticipated that the data will be voluminous, available on several different computers, and of varied quality. Because the data will come from different sources with varying degrees of quality, problems can be expected to arise in documenting the QA component of the data.

EMTS data sets would ideally all be transferred to the QA specialist at EMSL/LV for a QA check and annotation. However, the quantities of the data and the media used for storage at the NCC would present certain logistical difficulties for this transfer and often preclude the operation altogether. Nevertheless, quality assurance of the data included in the EMTS data base remains a primary goal; therefore, two methods are available for providing this validation.

1. If speed is a primary consideration, the QA check of the data will be carried out by the ERC QA specialist using dial-up access to the NCC IBM computer. The data set description, date of the check, and results of this check will be documented as a part of the normal QA function.

2. If the speed of the process is not critical, the QA process can be pursued entirely at the ERC location. A designated subset of the data set will be requested of the EMSL/RTP office and checked by ERC QA personnel. This subset will contain sufficient records for the QA check and will be extracted from the data set at as nearly random a method as practical. This would normally call for extracting one record out of each 100, 500, 1000, etc., records in the data set. The record extracted should occupy the same location in the sequence each time. For example, if a large data set is being examined, extracting the first record from the 28th record would cause the 128th, 228, 328th and 428th, etc., record to be extracted and placed in the set to be examined. If the resulting quantity of records is deemed insufficient, the process can be repeated to select more records from the data set, or the process can be modified to select more. Likewise, if too many records are selected, a similar process can be carried out to reduce the quantity of records to be examined, or the process may be repeated with more stringent criteria.

Digital Data--

EMSL/RTP will identify and locate as many data sets as possible that may contain data relevant to the site-characterization phase of the EMTS. These data may be from projects being run on the NCC computers or they may be from other sources. These other sources may be within EPA or other U.S. Government agencies, state or local governmental agencies, or nongovernmental agencies. The primary requirements at this phase are that the data be directly related to the six-county EMTS SMSA

geographic area and be generally less than five years old. There may be some exceptions to the time limit that will depend on uniqueness and quality of and the need for the data.

Data sets that are identified by EMSL/RTP for possible inclusion into the EMTS data base will be entered into the EDSIS, and a diskette of the description will be forwarded to the ERC. Because of the data flow from different sources and the constant updating of some of the data, continual updating of the EDSIS inventory is anticipated. ERC personnel will review the new data on the EDSIS and advise the EMSL/RTP office if the data set should be included in the EMTS data base.

The applicability of the data set to the goals of the EMTS is the responsibility of the ERC staff. Once data sets have been identified as thoroughly as possible, the description of the data set will be entered into the EDSIS system, and a diskette will be sent to ERC. ERC will evaluate the contents of the inventory and indicate that they feel EMSL/RTP should

1. obtain more information about the data set,

2. obtain the data set,

3. be prepared to obtain the data set, or

4. drop the data set from consideration.

If Option 2 is taken and the data must be purchased, the Project Coordinator will indicate to the EMSL/RTP office the funding source to be tasked for obtaining the data set. Costs incurred for Option 1 will be paid by EMSL/RTP.

EMSL/RTP will establish and maintain a comprehensive Master Index of all data sets in the EMTS data base maintained at the NCC. This Master Index will contain a narrative description of the data set, including the parameters it contains, the time frames covered, the locations or areas covered, and other pertinent data that may be used to adequately describe the data. The Master Index will be maintained on a computer file with one entry per data set in the EMTS data base maintained on the NCC computer. The Master Index will be a flexible document, and new data will be entered as they are developed. As more comprehensive understandings of the data sets are obtained, the Master Index will be updated. Because of these characteristics and the number of anticipated users, the use of printed versions of the index is discouraged. The file or portions of the file may be printed out for users who do not have access to an appropriate computer terminal; however, on-line access is preferred. In this way, all users will have the benefit of the most timely material.

EMSL/RTP staff must review the contents of each data set to define the entry in the Master Index. The staff will obtain data sets that have been approved for inclusion into the EMTS data base by the ERC and will review the data to fully characterize the data set, although some of this description may be available through the EDSIS. Supporting documentation will be reviewed if possible before the entry is created in the Master Index. Because the Master Index is a free-form description, a thorough understanding of the data is essential to entering accurate information.

Nondigital Data--

Some data may be identified in the site-characterization phase of the EMTS that are in a nondigital form. This will generally mean that the data set is printed on paper and is not available in a format that is machine readable. It may be either in a printed, tabular form (or already available on keypunch forms) or it may be in a form that requires interpretation to convert to digital formats (such as charts and graphs or maps). Some of these data will be most usable in the EMTS system in their nondigital format. These data resources will be cataloged into the EDSIS and retained in

original (or copy) form by ERC. So that EMTS resources are made available to the largest number of potential users, these documents will be stored and maintained by the central ERC office at Las Vegas.

Some nondigital data may be identified in the site-characterization phase that are useful to the EMTS only after the data have been digitized. As these data are identified, their description will be entered into the EDSIS system, and a diskette containing the entry will be sent to the ERC, just as the digitized data sets are handled. Upon receipt of the updated EDSIS inventory, the ERC staff will indicate to EMSL/RTP the action to be taken with the data. If the data set is to be entered into the EMTS data base, it must first be digitized. There are a variety of resources at both EMSL/RTP and EMSL/LV for converting nondigital data into digital form.

If the data do not require interpretation to be digitized, they may be entered into a data base or keypunched by EMSL/RTP or contractor personnel onto cards or magnetic tape for entry into the computer. In this case, a decision to perform the work will be made by the EMSL/RTP Project Officer based on the availability of personnel and of any necessary funding, the quantity of data to be converted into digital form, and the time available for the conversion.

If the data do require interpretation to be converted to machine-readable format, it must first be determined if suitably trained personnel are available to perform the interpretation without risking distortion of the data. This step is likely to be a major stumbling block and will probably prohibit the use of this type of data in the network. As a general rule, personnel with the expertise to interpret data are unskilled at data entry functions or they do not have the time required for the task of interpretation. Second, if personnel are technically capable of an accurate interpretation of the data and are available, suitable equipment for the interpretive effort must be located. Although varied types of digitizing tablets are available within the EMSL/RTP and EMSL/LV organizations, they may not be available for the EMTS project. Thirdly, the EMTS Project Officer must indicate a source of funding for the personnel and equipment involved in the conversion effort. Throughout this three-step determination, the EDSIS inventory system will continue to be updated with the status of the data set and its disposition.

EMTS Field Study Information

Field studies in the EMTS will in general rely on computers for processing data that are collected or developed during the studies because it is expected that large quantities of data will be collected, analyzed, and reported. It is not the intent of the EMTS project to provide computing capability to individual studies. The principal investigator is responsible for the data processing of his or her own data. Principal investigators within the EPA organization may elect to use the NCC and other computing resources available to them. In this regard, they must provide for the time-sharing funding necessary to process their own data. The EMTS system will not provide data management, data validation, or report preparation for individual studies.

Digital Data--

Field work data manipulations are generally handled best with microcomputers, whereas statistical or spatial analysis frequently requires minicomputers or mainframes. The results of these studies must be reviewed by the EMTS QA specialist and the determination of the QA of the data completed before the data are entered into the EMTS data base.

EMSL/RTP will not perform statistical analysis for principal investigators on the NCC IBM 3090-200 computer. The EMTS supports and facilitates statistical analysis for the principal investigator by collecting data from a variety of sources. The principal investigators may elect to use the computer

at the NCC for statistical analysis; however, they are responsible for providing their own time-sharing funding and are responsible for carrying out the analysis.

EMSL/RTP will assist principal investigators in converting digital data developed in EMTS projects into suitable formats and suitable media for inclusion in the EMTS data base. The staff will perform these conversions and will make the actual additions to the EMTS data base of the data after they are approved for inclusion and have passed the QA review. Supporting project documentation will be forwarded to the ERC for inclusion in the EMTS library. EMSL/RTP personnel will prepare the EDSIS entry and forward the diskette to the ERC. In addition, the data set developed for the EMTS project will be reviewed, and an entry describing the data set will be made in the Master Index maintained by EMSL/RTP.

Nondigital Data--

Data that have been developed by the principal investigators and that are not in digital format may be of significance to other projects. Thus, descriptions of these data are included in the EDSIS. However, EMSL/RTP will not normally digitize data generated from EMTS projects for inclusion into the data base. If other principal investigators need these data, they should reference the EDSIS and other documents maintained by the ERC. Principal investigators should present the ERC with a copy of their completed study data in digitized format for incorporation into the EMTS library for the use in other projects.

CARTOGRAPHIC DATA – EMSL/LV

EMSL/LV has the hardware, software, and trained personnel to use modern cartographic techniques for the display and spatial analysis of data sets that may be relevant to the EMTS. Spatial data are generally represented by mapping techniques that represent data at a single point in time over a geographic area. A Geographic Information System (GIS) will be used for the display and spatial analysis of data in the EMTS project.

Most of the spatial data sets that have been obtained for the EMTS will become part of the resources available to all users of the EMTS. Release of digitized data sets may not be possible because of licensing restrictions imposed by the sources of the data. But although the data may not be available to the principal investigators, maps created from these data can be used.

It is not the function of the EMSL/LV office to determine which parts of data may or may not be applicable to later studies. All spatial data sets that are accepted into and procured under the EMTS project will be stored at the EMSL/LV facilities for potential users. Under normal circumstances, these data will be available to any and all users of the system at no additional cost. (Any costs for extraction and processing must be borne by the principal investigators. The costs associated with access to the data on the DEC VAX computer are also the responsibility of the principal investigators.)

The data made available to principal investigators will be provided by EMSL/LV in the media and format requested through the Project Officer. The preferred media for the selected data should be discussed by the principal investigator and the EMSL/LV data management staff.

Attribute and map data sets that have been loaded into ARC/INFO become part of the GIS data base. Attribute data sets, will, for the most part, be derivatives of the data sets that have been collected and processed on the NCC IBM 3090-200 computer. The GIS data base has its own index system. An additional feature is a data dictionary that resides in INFO files on the VAX as part of the system; entries in the data dictionary are mapped one-to-one to files in the GIS data base. Entries to

the data dictionary are derived from information in the EDSIS and the Master Index maintained at the NCC, as well as new information as required to describe the unique nature of the ARC/INFO files.

Entries in the EDSIS will be cross-referenced to the data dictionary where possible to indicate whether those data sets have been used to generate a GIS data base file for EMTS. Thus, users can get a good idea of what is available in the GIS data base system by querying the EDSIS. When an actual data analysis is being planned, interaction with the GIS staff will be necessary. The EMSL/LV staff may assist users in the design of specific GIS programs to derive special map files and to customize data analysis as appropriate.

Site Characterization Information

During the site-characterization phase of the EMTS, spatial data will be evaluated, collected, stored, and processed on the DEC VAX 11/785 minicomputer by EMSL/LV personnel. The data collected in this phase of the project will consist of data sets that have been assembled for a primary purpose other than the EMTS project. Data sets that are exclusively spatial will be forwarded directly to the EMSL/LV facility for processing. Data sets that also have tabular components will, as a general rule, be processed first at EMSL/RTP and then forwarded to the EMSL/LV facility for processing of the spatial components.

It is anticipated that spatial data will be voluminous, available on a number of different computers, and of varied quality. Because the data come from different sources with varying degrees of quality, problems can be expected to arise in documenting the quality of the data.

Digital Data--

Data sets that have been identified by the EMSL-LV for possible inclusion into the EMTS will be entered into the EDSIS, and a diskette of the description will be forwarded to the ERC. Because of the data flow from different sources and the constant updating of some of the data, continued updating of the EDSIS inventory is anticipated. An updated copy of the EDSIS inventory will be provided by EMSL/LV to the ERC whenever it is updated. The ERC will, in turn, review the supplementary data and advise the EMSL/LV office of its preference of whether or not EMSL/LV should include the data set in the EMTS data base.

EMSL/LV and its subcontractors will identify and locate as many spatial data sets as possible that may contain data relevant to the site-characterization phase of the EMTS. These data may be from projects being run on the EPA computers or they may be from other sources. These other sources may be U.S. Government agencies, state or local governmental agencies, or nongovernmental agencies. The primary requirements at this phase are that the data be directly related to the six-county EMTS SMSA geographic area and optimally be less than five years old. It is anticipated that there will be exceptions to the time limit based on the need for and the quality and currency of the data.

The applicability of the data set to the goals of the EMTS is the responsibility of the ERC staff. Spatial data sets will be treated the same as tabular data sets in that a description of the data set will be entered into the EDSIS and the ERC staff will evaluate the data set to determine future actions.

Nondigital Data--

Some data sets may be identified in the site-characterization phase of the EMTS that are in a nondigital form. This will generally mean that the data set is printed on paper and is not available in a format that is machine readable. Spatial data may be either in a printed, tabular format (or

already available on keypunch forms), or they may be in a form that requires interpretation to convert to digital formats (such as charts and graphs or maps).

Some of these data may be more usable in the EMTS system in a nondigital format. These data resources will be cataloged into the EDSIS and retained in original (or copy) form. So that EMTS resources are made available to the largest number of potential users, these documents will be stored and maintained in the central EMTS library at EMSL/LV maintained by ERC.

Some nondigital data may be identified in the site-characterization phase that are useful to the EMTS only after they have been digitized. As these data are identified, they will be entered into the EDSIS, just as digitized data sets are. Upon receipt of the updated EDSIS inventory, the ERC staff will indicate to EMSL/LV the action to be taken with the data. If the data set is to be entered into the EMTS data base, it must first be digitized. There are a variety of resources at EMSL/LV for converting nondigital data into digital form.

If the data do not require interpretation to be digitized, they may be entered into a data base by EMSL/LV or digitized by contractor personnel onto magnetic tape for entry into the computer. In this case, a decision to perform the work will be made by the EMSL/LV Project Officer based on the availability of personnel and of any necessary funding, the quantity of data to be converted into digital form, and the time available for the conversion.

If the data do require interpretation to be converted to machine-readable format, suitably trained personnel must be available to perform the interpretation. Second, suitable equipment for the interpretive effort must be located. Third, the EMTS Project Officer must indicate a source of funding for the personnel and equipment involved in the conversion effort. All changes in status of the data set will be noted in the EDSIS.

EMTS Field Study Information

Field studies in the EMTS will, to a large extent, rely on computers for the processing of data that are collected or developed during the study because it is expected that large quantities of data will be processed. This collection and analysis function is greatly enhanced with the power of modern computers. Microcomputers are suitable for handling field work data, whereas minicomputers or mainframes are frequently required for statistical or spatial analysis.

It is not the intent of the EMTS project to provide computing capability to projects. The principal investigator must make his own arrangements for his study and provide funding for the processing of his data. Principal investigators within the EPA organization may elect to use the EPA computers and any other computing resources available to them. In this regard, they must provide for the funding necessary to enter and process their data. The EMTS system will not provide personnel to enter data, validate data or create reports for individual studies.

If the principal investigators elect to utilize the GIS resources available within the EMTS for their project needs, the EMSL/LV staff can provide them access to site-characterization data and data from other studies that have produced completed data. This access to data may be a significant benefit to the investigator in researching the site-characterization data.

Digital Data--

Data developed by the principal investigators for their projects should be included in the EMTS data base for use in other projects. However, the results of these studies must be reviewed by the EMTS QA specialist and the determination of the QA of the data completed before the data are entered into the EMTS GIS data base.

In most cases, EMSL/LV will not perform spatial analysis for principal investigators on the EMSL/LV DEC VAX 11/785 computer. The EMTS makes the resources available to the principal investigator and will, in some ways, facilitate the spatial analysis of data. The principal investigators may elect to use the resources of the EMTS, but they must provide their own funding and are responsible for carrying out the analysis.

EMSL/LV will assist principal investigators in converting digital data developed in EMTS projects into suitable formats and suitable media for inclusion in the GIS data base. The staff will make the actual additions to the EMTS data base after the data have been approved for inclusion and passed the QA review. Supporting documentation will be included in the EMTS library. EMSL/LV personnel will prepare the EDSIS entry.

Nondigital Data--

Data that have been developed by the principal investigators that are not in digital format may be of significance to other projects. Thus, descriptions of these data are included in the EDSIS. However, EMSL/LV will not normally digitize data from EMTS projects for inclusion into the data base. If other principal investigators need these nondigital data, they must refer to the hard-copy documents that are maintained by the ERC and may elect to digitize the data. Principal investigators should present the ERC with a copy of their completed study for incorporation into the EMTS library for the use in other projects.

DATA FLOW (COMMUNICATIONS)

The geographic separation of the two laboratories' computer centers involved in the EMTS creates some potential problems for the smooth, timely, and accurate transmission of data between them. Collecting and processing the data in a coherent, efficient manner and making them available to the EMTS principal investigators from the two sites require extensive planning and coordination.

The principal flow of data is expected to be from the EMSL/RTP location to the EMSL/LV location. All data sets included in the EMTS data base except those that are exclusively spatial representations will be first entered into the EMTS system through the data base set up and maintained at the NCC by EMSL/RTP. Data sets that are exclusively spatial and that may be represented only with the GIS at EMSL/LV will be processed and maintained on the VAX at the EMSL/LV location.

It is anticipated that users will request customized extractions of the data base, selecting and comparing only those portions of the data base that are of significance for a particular study. EMSL/RTP will create the requested customized data set for the user. If the user wishes to pursue only a statistical analysis of the selected data set, and he elects to do so on the NCC IBM 3090-200 computer, transfer of the data to EMSL/LV will not be required.

If the user wishes to also utilize the GIS for analysis of his data, the selected data set must be transferred to the DEC VAX 11/785 computer at the EMSL/LV location. The EMTS staff will transfer data sets from the EMSL/RTP location to the EMSL/LV location principally by magnetic tapes.

For the EMTS there are three primary data communication needs for the data that will constitute the EMTS data base.

1. Flow of data into the system, in which the principal flow of data will be into the NCC data base, and data that are exclusively spatial will flow directly to EMSL/LV for inclusion in the GIS.

2. Flow of data from the NCC IBM 3090-200 computer to the EMSL/LV GIS system.

3. Flow of data from both of these operations to the ultimate users of the system (the principal investigators).

Each of these data flows requires controls and management procedures to ensure that the needs of the users are met with a single, comprehensive approach. The management and control structure of the system places additional restraints on the system because of the different organizational structures and varied EPA–contractor relationships. The system of inventories, indexes, and catalogs in the EMTS system is in place to provide as much assistance as practical to meet the users needs.

Flow of Data into the EMTS Data Base

The first flow, that of data into the EMTS data base, can come from two types of sources. Data that are gathered for some primary purpose other than an EMTS project are part of the site-characterization phase. Data that are gathered directly for an EMTS project are part of the project phase and are also to be added to the data base. The primary difference between these two types of data is in who handles the data. The EMTS staff will locate, process, and add the site-characterization data to the data base. The project data will be collected, processed, and documented by the principal investigators before the data are added to the data base.

Digital data (data that are in a format that is machine readable without keypunching and/or digitizing) will be transferred into the EMTS data base principally through magnetic tapes. Both the NCC IBM and the EMSL/LV DEC VAX computers have the capability to read these tapes, they are relatively inexpensive, they are accurate, and they form a permanent archive. Data are very rarely lost in transmission when this method is used. Most computer installations that will be creating tapes of data are able to create the tapes with formats that are readable on the IBM or VAX computers, as appropriate. In addition, these computers can convert data from other formats on a case-by-case basis, as necessary.

Nondigital data (data that must be converted from their present format to be read by the computer) may, under certain conditions, be entered into the EMTS data base. It is necessary for the Project Officer to indicate the usefulness of the data set to other studies in the EMTS and have an economic justification for this to happen.

Site-characterization data may be gathered into the NCC IBM computer for storage, processing, and possible later transfer to the GIS system; or, if the data are exclusively spatial, they may be entered directly into the GIS system without going through the NCC. In either case, a description of the data set must be entered into the EDSIS, and QA of the data set must be documented. Field study data that are developed by principal investigators may also be handled this way.

The EMTS staff will provide blank data tapes to facilitate the transfer of data if the owner of the data requests them. The tapes will be standard, ½-in., 6250-BPI tapes. The computer center at which the data are requested will supply the tapes.

Data sets may be transferred into the EMTS via other means if necessary. In those cases where time is a factor, data may be transferred into the EMTS data base directly through an on-line connection with dial-up telephone lines. However, this method poses the added risk of lost data, and the cost of the operation is generally higher than that incurred by the use of tapes.

Flow of Data from the NCC to the GIS System

The second major flow of data in the EMTS is the flow of selected data sets from the NCC computer at RTP to the GIS computer in Las Vegas. The normal and preferred method of transferring these data is via magnetic tapes. Standard ½-in., 6250-BPI computer tapes will be unlabeled and prepared on the NCC IBM computer. They will be transferred to the operations section of the GIS computer staff by using an overnight express carrier. The documentation accompanying each tape will provide the GIS staff a count of the records on the tape, the record lengths, the blocking factor, a record description, and any other data necessary to facilitate the processing of the data.

A less desirable method of transferring data to the GIS from the NCC is by use of the EPA data network. This method of data transfer should be attempted only in those situations where immediate transfer of the data is essential. The EMTS staff at EMSL/RTP will prepare the data set and then transfer it to EMSL/LV via the network. The terminating site at Las Vegas is a VAX computer; however, this is not the same VAX that runs the GIS. There is a link between the two VAX computers, which should be used to communicate with the GIS. Also, a magnetic tape can be created on either of the VAX computers for use in the GIS. It is necessary for the EMSL/RTP staff to inform the EMSL/LV staff of the record size, blocking factor, number of records being transferred, record description, etc., to ensure the completeness and integrity of any data set transferred across the country.

Flow of Data from the EMTS Data Base to the User

The third data transfer mechanism in the EMTS is the transfer of data from the EMTS data base to the user community. This process is complicated by the geographic distance separating the two principal locations where the data may be stored. The single organizational entity that has knowledge of the data stored in both systems is the ERC at the University of Nevada at Las Vegas. Requests for data should normally be made by the principal investigator through the Project Coordinator to the ERC.

It is permissible for the principal investigator to directly access the EMTS resources without going through the staffs at EMSL/RTP or EMSL/LV. However, it is essential that the principal investigator have the technical knowledge and have met the administrative requirements for access to the resources. In addition to satisfying the requirements for access to the computer systems, establishing accounts, etc., it is also necessary that a user of any component of the EMTS system be identified to the party responsible for the EMTS data base. For example, a user may not access the EMTS GIS unless he has been granted that access by EMSL/RTP personnel responsible for the integrity and use of the system.

The best place to extract the data from the EMTS data base will be determined at that time and will depend on the type of data desired and the portrayal method. Spatial data will be best provided by the EMSL/LV location, and tabular data sets will best be created by the NCC in RTP.

The ERC will communicate the request for data to the appropriate agency (in some cases, a single request could result in queries being made against both data bases) and provide the name and contact point of the requester. When the request has been filled and the data have been forwarded to the principal investigator, the staff at the EMSL/RTP and/or EMSL/LV will inform the ERC so the request records may be closed.

Data will generally be provided to the requester in one of three ways. Printed output of the data requested will probably be used most often for short retrievals. The requested data may be put on magnetic tape or on an IBM standard 5¼-in. diskette for more lengthy retrievals or when requested. For some applications and with experienced users, it may be appropriate for the output to be presented as a data set on one or more of the computers used in the EMTS. This form would

most likely be used when the principal investigator wished to conduct an analysis of the data on the computer.

This third major flow of data may be divided into two parts, the flow of data that were gathered during the site-characterization phase of the project and the flow of data that have been gathered as a part of the specific EMTS project under the supervision of the principal investigator. Procedures for the first flow of data are straightforward and are as outlined in this section. However, the flow of EMTS project data from the data base to another principal investigator could raise certain procedural problems.

The request for data from other EMTS projects will be forwarded through the EMTS Steering Committee by the ERC staff. This committee will determine the applicability of the data that were previously gathered to the purpose stated in the request. If the EMTS Steering Committee approves the request, the project officer who collected the data will be asked to share his data with the requester. The data will not be released by the EMTS staff to other EMTS projects until the concurrence of the originating project officer has been received.

DATA SECURITY

EMTS data stored in EPA computer systems must be protected from unintentional or unauthorized access to prevent erroneous modifications, accidental erasure, or premature removal from the system. Three types of safeguards have been established to ensure this protection: access control, system and data set backups, and on-line documentation of data sets in EMTS libraries.

Access Control

Access to the computer system is controlled by passwords. Also, accounting data with currently funded projects must be entered to permit access to the system. Additionally, the data base management systems at NCC and EMSL/LV each have a complex, multilevel password system built into them. Finally, the systems provide a tracer or log of access to data sets that can be researched if unauthorized access to the data is suspected.

Backups

The data sets will be stored in both locations in a modern, fourth-generation data base management system, which will provide security and backup procedures as a regular part of the maintenance of the data. Additional backups of the EMTS data sets on magnetic tapes will be made by the EMTS staff to ensure against accidental erasure, and to ensure that the data are not removed because they have not been accessed over a period of time. These system tapes will be maintained in the tape library with an indefinite expiration date. In this way, EMTS data sets can always be recovered.

ERC Library Documentation

The ERC is responsible for maintaining libraries of data to support the EMTS. Among the data stored are descriptions of data sets, maps of data stored in the ISDAS, procedures used for statistical manipulations, and documentation referring to the QA of the EMTS data sets. Also available is supporting documentation for data sets that have been procured for use in the EMTS. The results of EMTS studies should be included in this documentation. The reader is referred to the ERC for a more detailed listing and comprehensive description of the documentation maintained in the ERC libraries.

REFERENCES

1. Environmental Research Center. Environmental Methods Testing Site Users Guide (Undated Draft). Environmental Research Center, University of Nevada – Las Vegas, Las Vegas, NV.

2. Environmental Research Center. EMTS Quality Assurance Program Plan (Draft). Environmental Research Center, University of Nevada – Las Vegas, Las Vegas, NV, 1986.

3. Environmental Research Center. Environmental Methods Testing Site Project Management Plan, February 6, 1986. Environmental Research Center, University of Nevada – Las Vegas, Las Vegas, NV, 1986.

4. U.S. Environmental Protection Agency. NCC-IBM Users Guide (Revised May 1987). National Data Processing Division, U.S. Environmental Protection Agency Office of Administration and Resources Management, Research Triangle Park, NC, 1987.

APPENDIX A

Figure A-1. Detailed diagram of the overall NCC communications network.

Figure A-2. Detailed diagram of the Region 4 communications network.

Figure A-3. Detailed diagram of the EMSL/LV communications network.

Figure A-4. Detailed schematic of the NCC IBM 3090-200 mainframe computer.

Figure A-5. Detailed schematic of the ISDAS.

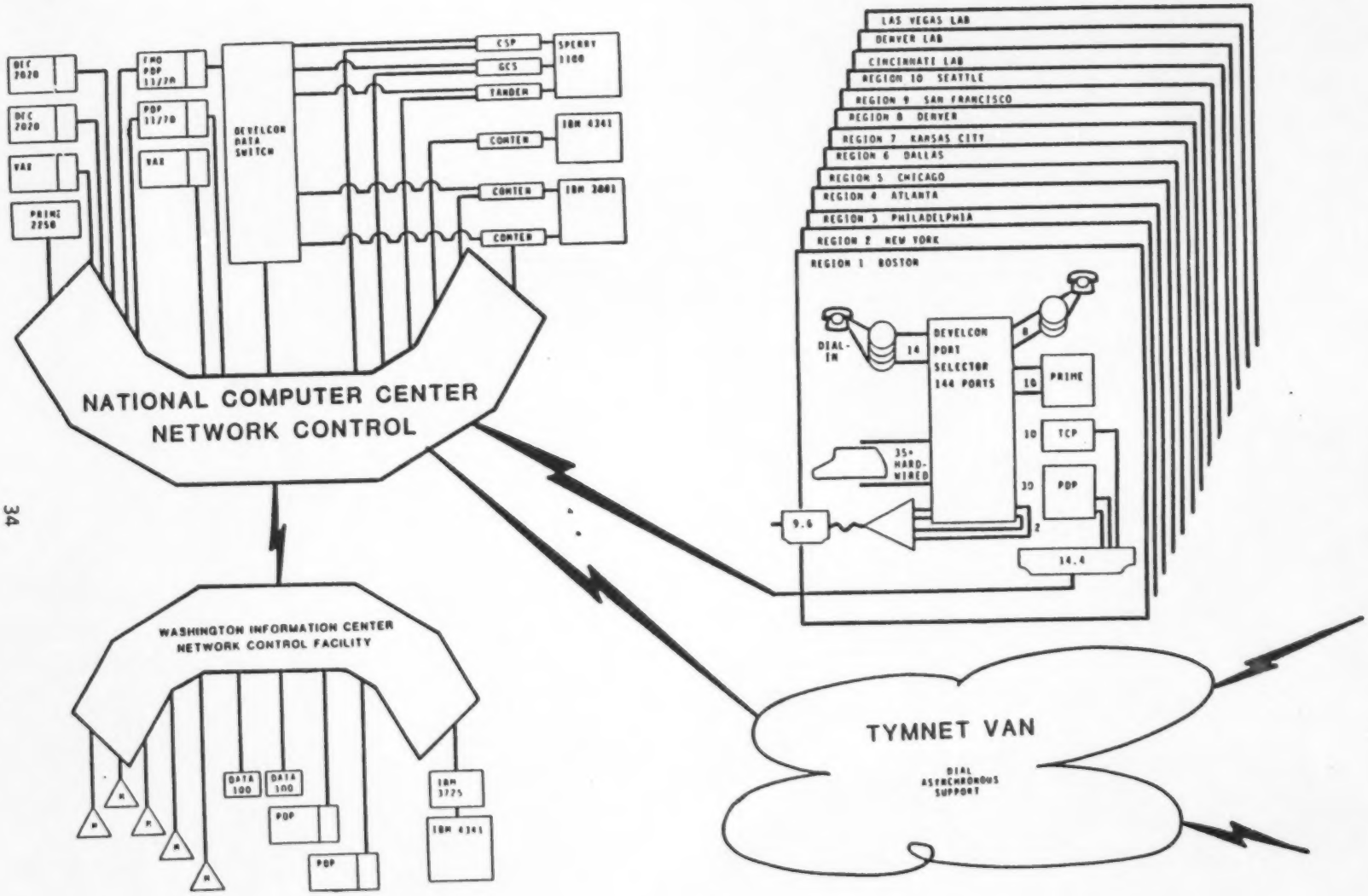

Figure A-1. Detailed diagram of the overall NCC communications network.

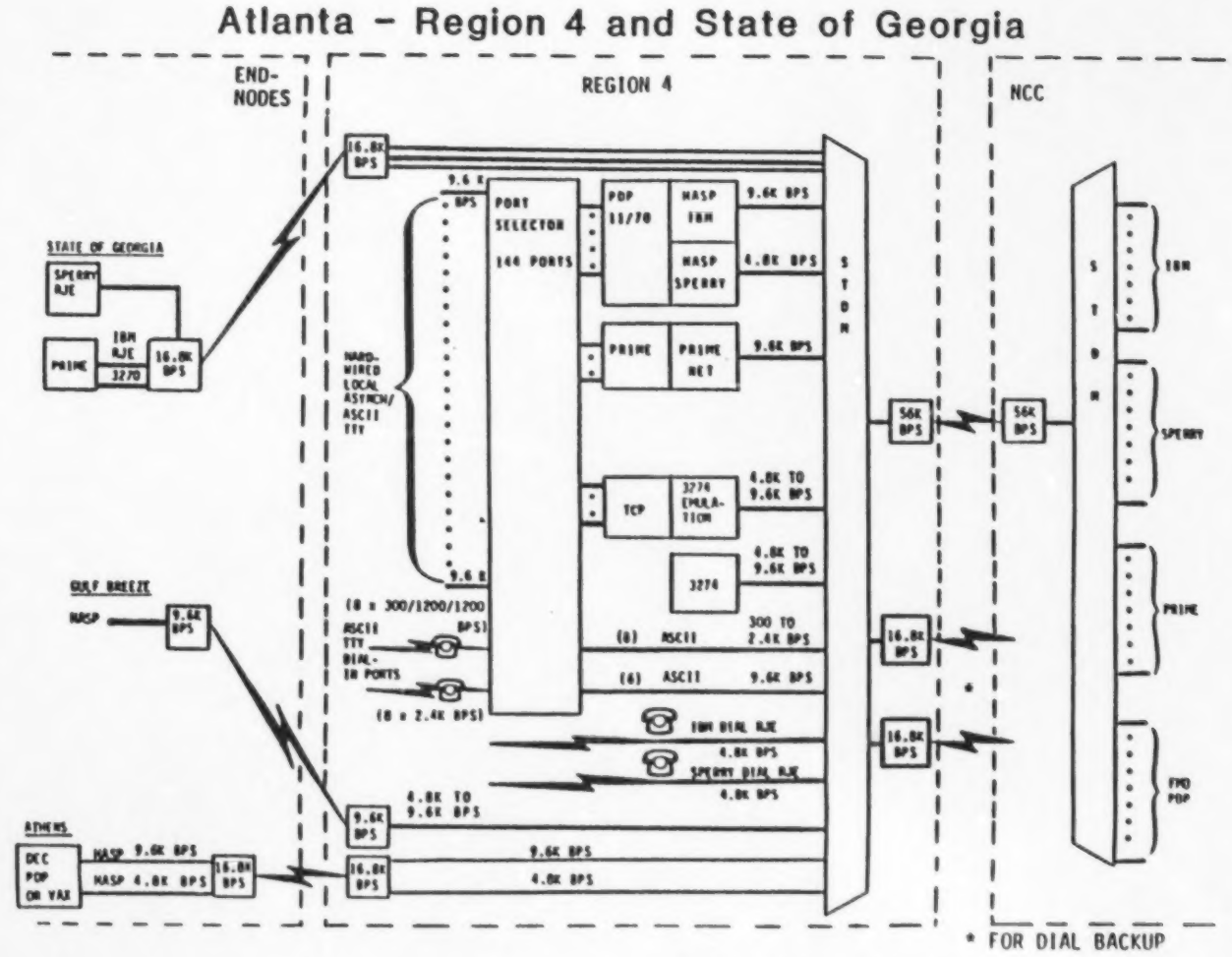

Figure A-2. Detailed diagram of the Region 4 communications network.

Reproduced from best available copy.

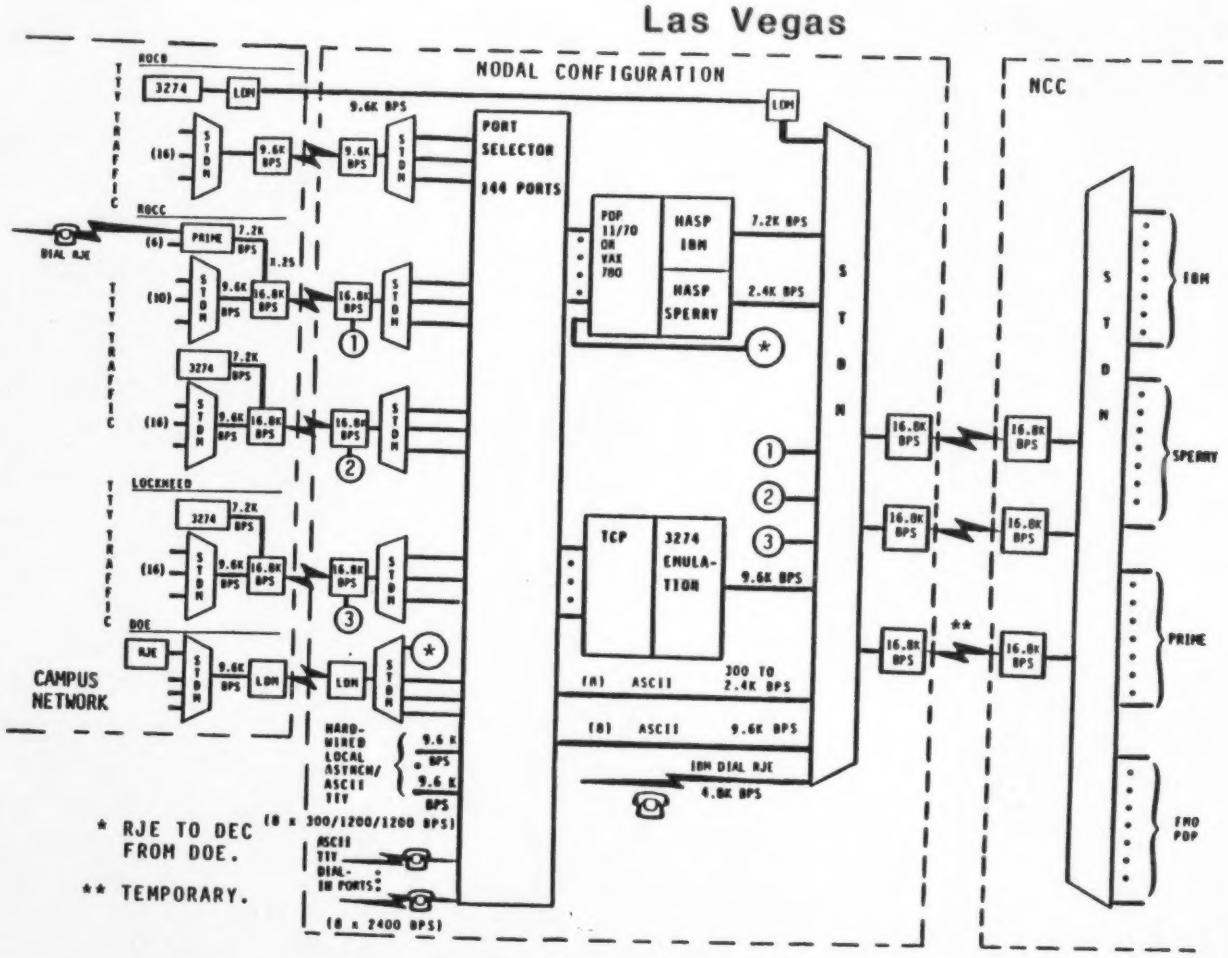

Figure A-3. Detailed diagram of the EMSL/LV communications network.

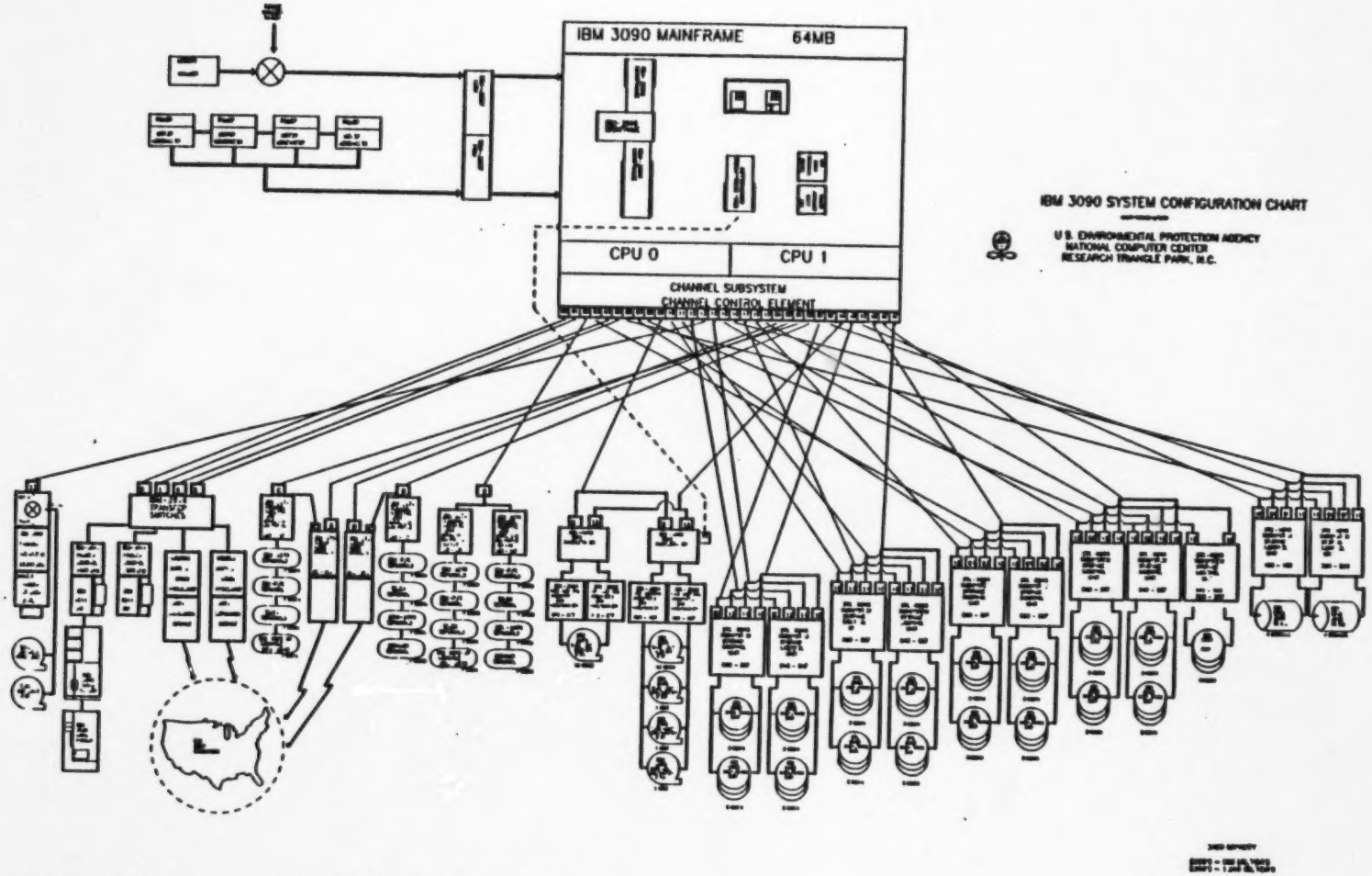

Figure A-4. Detailed schematic of the NCC IBM 3090-200 mainframe computer.

Reproduced from
best available copy.

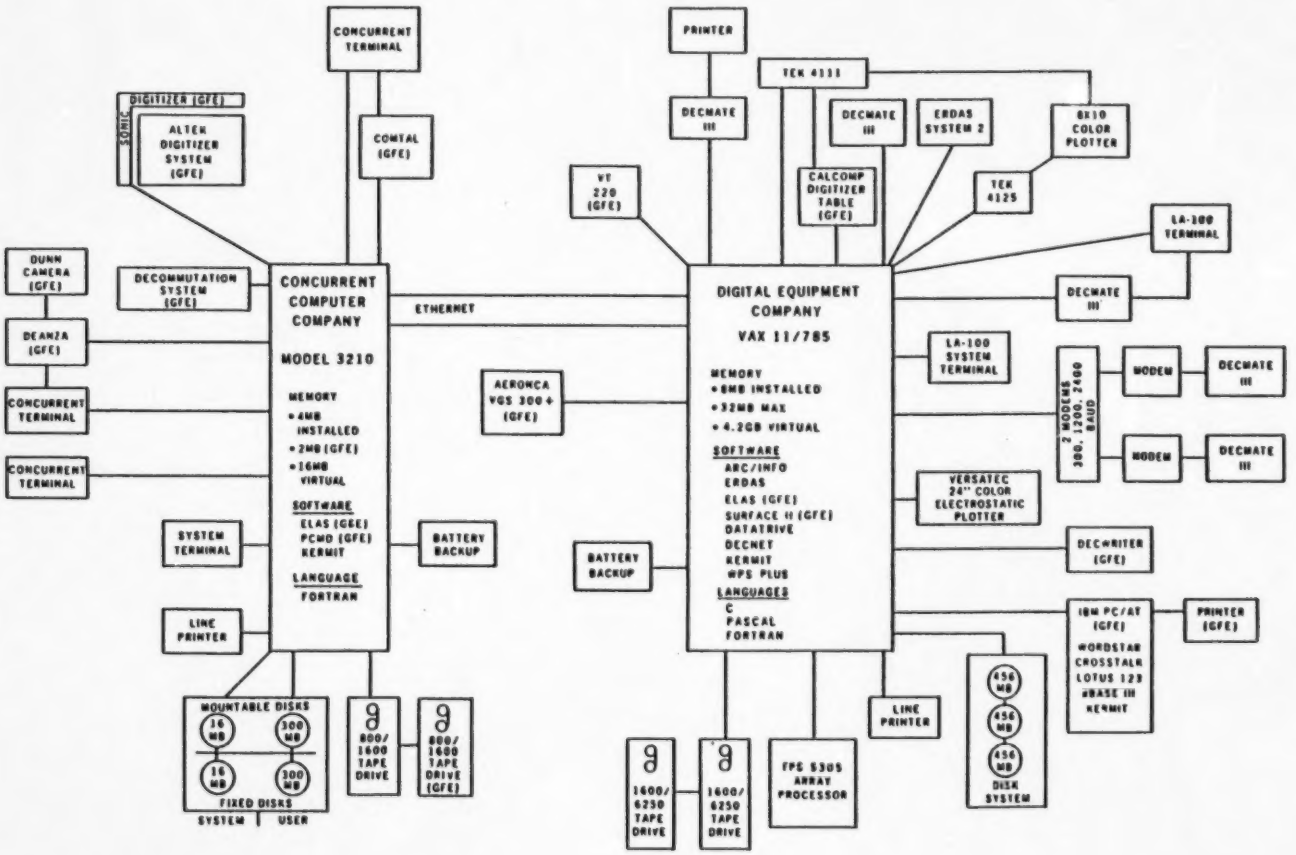

Figure A-5. Detailed schematic of the ISDAS.